Universität Hildesheim
Informationswissenschaft

Thomas Mandl
Christa Womser-Hacker
(Hg.)

Effektive Information Retrieval Verfahren in Theorie und Praxis:

Proceedings Fünfter Hildesheimer Evaluierungs- und Retrieval (HIER) Workshop

11.10.2006
Universität Hildesheim

Inhalt

Vorwort

Dieser Band fasst die Vorträge des Fünften Hildesheimer Evaluierungs- und Retrieval-Workshops (HIER) zusammen, der am 11. Oktober 2006 an der Universität Hildesheim stattfand. Die HIER Workshop-Reihe begann im Jahr 2001 mit dem Ziel, die Forschungsergebnisse der Hildesheimer Informationswissenschaft zu präsentieren und zu diskutieren. Mittlerweile nehmen immer wieder Kooperationspartner von anderen Institutionen teil, was wir sehr begrüssen. Alle Beiträge stehen dieses Jahr in enger Beziehung zu den Kernthemen Information Retrieval und Evaluierung. Traditionell bietet der HIER auch ein Forum für Systemvorstellungen und praxisorientierte Beiträge.

In diesem Jahr schliesst sich der HIER an den Workshop Information Retrieval 2006 an, den die Fachgruppe Information Retrieval der Gesellschaft für Informatik (GI) veranstaltet hat. Dieser fand erstmals gemeinsam mit den Tagungen der Fachgruppen *Adaptivität*, *Maschinelles Lernen* und *Wissensmanagement* im Rahmen der Workshop-Woche *Lernen - Wissensentdeckung - Adaptivität* (LWA, 09.-11.10.2006 an der Universität Hildesheim) statt. Die Veranstaltung setzte die erfolgreiche Reihe von Workshops der Fachgruppe fort (http://www.fg-ir.de).

Unser Dank gilt diesmal den Organisatoren der LWA, welche den HIER in diesen Rahmen einbetteten, so Interessenten beide Events wahrnehmen konnten, ohne dass zusätzlicher Aufwand entstand. Insbesondere danken wir Herrn Alexandre Hanft, Martin Schaaf und Prof. Dr. Klaus-Dieter Althoff. Weiterhin danken wir Herrn Benjamin Ahlborn von der Universiätsbibliothek Hildesheim herzlich, der den Tagungsband auf dem Medienserver der Universitätsbibliothek online zur Verfügung stellt.

Zuletzt noch ein Blick in die Zukunft: Auch 2007 findet in Hildesheim wieder ein Workshop statt, diesmal zum Thema *Mehrsprachigkeit in Informationssystemen*. Den organisatorischen Rahmen bietet die Jahrestagung der Gesellschaft für Angewandte Linguistik (GAL), die im September 2007 an der Universität Hildesheim stattfinden wird.

Wir hoffen auf große Resonanz und freuen uns auf den wissenschaftlichen Austausch der bis dahin gewonnenen Erkenntnisse.

Hildesheim, Oktober 2006

Christa Womser-Hacker, Thomas Mandl

Methodische Aspekte der Evaluation kooperativer E-Learning-Szenarien

Joachim Griesbaum

Universität Konstanz
Informationswissenschaft
Fach D 87
D-78457 Konstanz
griesbau@inf.uni-konstanz.de

Zusammenfassung

K3 ist ein Forschungsprojekt, welches das Ziel verfolgt die distributiven und kommunikativen Mehrwertpotenziale asynchroner Medien Gewinn bringend für die universitäre Ausbildung zu nutzen. Hierzu werden aufsetzend auf dem von Kuhlen vorgeschlagenen Paradigma des netzwerkbasierten Wissensmanagements konzeptionelle didaktische Ansätze erprobt und eine kollaborative Wissensmanagementsoftware entwickelt. Dieser Artikel beschreibt zunächst den grundlegenden Ansatz und wichtige Gestaltungsfaktoren des netzwerkbasierten Wissensmanagements. Darauf aufbauend werden methodische Aspekte der Evaluation solcher kooperativer Lernszenarien dargestellt, Untersuchungsinstrumente angeführt und die Reichweite und Grenzen der Evaluierbarkeit derartiger Lernszenarien diskutiert.

1 K3 – netzwerkbasiertes Wissensmanagement

K3 steht für Kollaboration, Kommunikation und Kompetenz und ist ein Forschungsprojekt, welches die lernförderliche Umsetzung des kollaborativen oder auch netzwerkbasierten Wissensmanagements in der universitären Ausbildung anvisiert[1]. Kollaboratives Wissensmanagement basiert auf der Idee, die Potenziale netzbasierter Wissenskommunikation, Wissensgenerierung und Wissensnutzung für das individuelle und gruppenbezogene Lernen zu nutzen, indem asynchrone Medien, zuvorderst Kommunikationsforen dazu verwendet werden, um wechselseitigen Austausch und Kooperation zwischen den Teilnehmern eines Kurses zu befördern [Griesbaum 2006], S. 200.

Das K3-Projekt ist sowohl als empirisches Feldprojekt zur Erprobung von Konzepten des netzbasierten kooperativen Lernens einzuordnen als auch als technologisches Entwicklungsprojekt zu sehen, in dem zugleich eine forenbasierte kollaborative Lernumgebung

[1] K3 wird an der Universität Konstanz am Lehrstuhl Informationswissenschaft (Prof. Kuhlen) entwickelt. Es handelt sich dabei um ein vom BMBF (DLR PT-NMB+F) im Rahmen des Programms „Innovation und Arbeitsplätze in der Informationsgesellschaft des 21. Jahrhunderts“ in Bezug auf die Fachinformation gefördertes Projekt (Projektnummer: 08C5896). Weitere Informationen unter http://www.k3forum.net.

entwickelt wird [Kuhlen 2002]. Hierzu werden traditionelle Lernmethoden aus Face-to-Face-Szenarien mit netzbasierten Wissen generierenden Lernmethoden "angereichert", ein neues Leistungsbewertungssystem genutzt und eine Wissensmanagementsoftware entwickelt, welche eine Vielzahl von Technologien zur Unterstützung von Wissenskommunikation und Wissensgenerierung zur Verfügung stellt.

2 Erfolgsfaktoren des netzwerkbasierten Wissensmanagements

Für die Ausgestaltung des kollaborativen Wissensmanagements in Hochschulkursen existieren keine allgemeingültigen Rezepte. Angesichts komplexer Zusammenhänge zwischen den Eigenschaften einzelner Teilnehmer (etwa Medienkompetenz, Vorwissen, Motivation), der Lerngruppen (z.B. Wissensverteilung, Klima, Kohäsion), Lernumgebung (Curriculare Integration, Didaktisches Design und Technologie) ist es einleuchtend, dass die Ergebnisse vom Zusammenwirken multipler, interdependenter Wirkungsflüsse abhängig sind [Friedrich & Hesse 2001].

Als Gestaltungsfaktoren des netzwerkbasierten Wissensmanagements lassen sich primär die angeführten Eigenschaften bzw. Inputfaktoren der Lernumgebung anführen. K3 setzt dabei auf Konzepte und Technologien, die zunächst grundsätzlich auf die erfolgreiche Bewältigung der Anfangssituation und die dauerhafte Aufrechterhaltung der Motivation angelegt sind, weitergehend eine, aus didaktischer Perspektive, lernförderliche inhaltliche und organisatorische Ausgestaltung der kooperativen Lernprozesse anvisieren und schließlich darauf abzielen, auf technologischer Ebene im Vergleich zu einer auf Standardtechnologien basierenden Umsetzung durch die Bereitstellung direkt am Lernprozess orientierter „Lerntechnologien“ die Reichweite und Effektivität der Werkzeugunterstützung zu erhöhen [Griesbaum 2006], S.163-166.

3 Methodische Aspekte der Evaluation kooperativer Lernszenarien

Auf welche Weise lässt sich überprüfen, ob und inwieweit sich die postulierten Mehrwerte des netzwerkbasierten Wissensmanagements in realen Hochschulkursen tatsächlich realisieren (lassen)? Wie lässt sich feststellen, welche curricularen, didaktischen und technologischen Unterstützungselemente des kooperativen E-Learning[2] sich in ihrem kombinatorischen Zusammenwirken wie auswirken?

Aufgrund der Komplexität der Wirkungsflüsse derartiger Lernszenarien sind paradigmenbasierte Evaluationsstandards nicht vorhanden bzw. kaum denkbar [Pfister 2004], S. 5. Hinsichtlich einer Ergebnisanalyse ist der erzielte Lernerfolg von zentralem Interesse. Dabei ist sowohl der individuelle als auch der kooperative Lernerfolg zu überprüfen, ein sehr schwieriges Unterfangen [Wessner et al. 1999]. Daneben gilt es, Kostenaspekte wie den zeitlichen oder technologischen Aufwand zu berücksichtigen. Weiterhin sollte Evaluation Lernprozess begleitend versuchen, die Bedingungen und Prozesse erfolg-

[2] Die Begrifflichkeiten netzwerkbasiertes oder kollaboratives bzw. kooperatives Wissensmanagement sowie CSCL und kooperatives E-Learning werden im vorliegenden Text vereinfachend, soweit nicht anders angegeben synonym behandelt. Zu den wesentlichen Unterschiede und Bedeutungsnuancen der Termini vgl. [Griesbaum 2006].

reichen Lernens nachzuvollziehen und auf diese Weise Erfolgsfaktoren netzbasierten kooperativen Lernens identifizieren. Letzten Endes liegt das Ziel von Evaluationen nicht nur darin, deskriptiv gültige Zusammenhänge aufzudecken, sondern Prozess begleitend praxistaugliche Ergebnisse und Verfahren zu erschließen. Evaluation ist also gerade im hochkomplexen Themenfeld des netzbasierten kooperativen Lernens nicht nur theoriegeleitet, sondern auch in starkem Maße anwendungsorientiert [Reinmann-Rothmeier et al. 2001], S. 132-133.

4 Untersuchungsmethoden

Grundsätzlich stehen für die Evaluation des netzwerkbasierten Wissensmanagements alle Erhebungs- und Auswertungsmethoden der qualitativen und quantitativen Sozialforschung wie Befragungen, Beobachtungen, Tests, Dokumentanalysen zur Verfügung [Schwarz 2001]. Der Entwurf bzw. die Konzeption eines konkreten Forschungsdesigns ist allerdings sehr anspruchsvoll und problembehaftet [Pfister 2004], S. 12.

[Haake et al. 2004b] [Wessner et al. 1999] plädieren bei der Neuentwicklung kooperativer netzbasierter Lernumgebungen für eine verschränkte formative Evaluation auf zwei Ebenen: Zum einen auf der Ebene der Entwicklung der Lernumgebung, zum anderen auf der Ebene der Evaluationskriterien selbst. Dem gemäß werden die Ergebnisse der Evaluation einerseits zur Optimierung des Lernszenarios genutzt, zum anderen werden die Evaluationskriterien im Ablauf selbst, gemäß den neuen Erkenntnissen und Randbedingungen, die erst im Entwicklungsprozess sichtbar werden, modifiziert. Diese Verschränkung verdeutlicht die Komplexität von Evaluationen und stellt klar, dass bei der Evaluation neuer, in realen Lernkontexten entwickelter kooperativer Lernumgebungen der Evaluationsgegenstand selbst fortlaufend Veränderungs- und Entwicklungsprozessen unterliegt und damit quasi nie ein identisches Treatment darstellt [Schwarz 2001]. Setzt man diesen Sachverhalt mit den oben dargestellten Wirkungsflüssen in Beziehung, so ist zu erwarten, dass es sehr schwierig ist, Evaluationen gültig (valide), zuverlässig (reliabel) und verallgemeinerungsfähig (generalisierbar) [Schnell et al. 1999], S. 145-160 auszugestalten.

Genau dies sind die Gütekriterien „objektiver“ quantitativer Untersuchungsmethoden. Diese zielen darauf, quantifizierbare Sachverhalte möglichst standardisiert zu erheben und basierend auf statistisch fundierten Auswertungen valide Aussagen über kausale bzw. korrelative Beziehungen im Sinne von Ursachen-Wirkungszusammenhängen abzuleiten [Pfister 2004]. Folgende Abbildung veranschaulicht ein grundlegendes experimentelles Design.

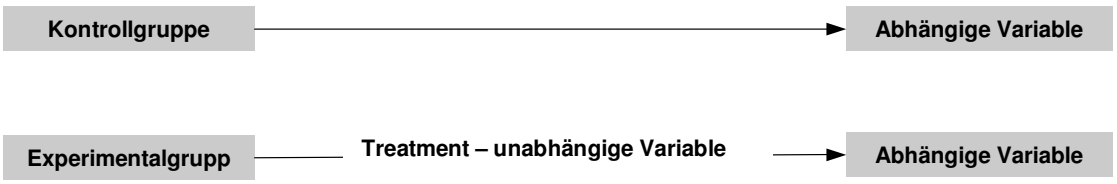

Abbildung 1: Grundlegendes experimentelles Design (in Anlehnung an [Pfister 2004], S. 7)

ihre Grenzen bzw. sind kaum zu realisieren [Dyson & Campello 2003]. [Schulmeister 2002], S. 398 konstatiert generell für die Methodologie von Vergleichsevaluationen im E-Learning die weit verbreitete Problematik nicht kontrollierter bzw. auch kaum zu kontrollierender Variablen. Bei der Anwendung bzw. Übertragung solcher Verfahren in Feldszenarien lässt sich zudem aus ethischer Perspektive die Frage stellen, ob und inwieweit es legitim ist, zum Zwecke einer höheren Kontrolle Kontrollgruppen zu bilden und damit Lernende verschiedenen Treatments auszusetzen, denen unterschiedliche Lernchancen bzw. Erfolgsaussichten zugeordnet werden [Dyson & Campello 2003], S. 12.

Neben diesen grundsätzlichen ethischen bzw. Komplexitätsproblemen stellt sich insbesondere im Bereich der Prozessevaluation, also der Analyse der kooperativen Lernprozesse, die Frage nach der substanziellen Adäquanz rein quantitativer Methoden. Es ist fraglich, inwieweit quantifizierbare Diskursstrukturparameter bzgl. einer Beurteilung der Qualität, im Sinne einer lernförderlichen Güte des Ablaufs der Lernprozesse, sinnvoll operationalisiert werden können. Für die lernförderliche Qualität des Diskurses sind weniger statistisch operationalisierbare Struktur- oder Verlaufsvariablen einschlägig, vielmehr ist die Güte der Diskussion auf semantischer Ebene relevant. Um diese semantische Dimension kooperativer Lernprozesse besser zu erfassen, können inhaltsanalytische Verfahren genutzt werden [Schnurer 2005], S. 59.

Bei derartigen Verfahren steht die inhaltliche Interpretation von individuellen und sozialen Handlungen im Fokus. Während quantitative Methoden eher darauf zielen, vorgegebene Fragestellungen bzw. Hypothesen zu erschließen, lassen sich solche qualitative Verfahren vor allem auch dazu nutzen, unbekannte Zusammenhänge aufzudecken [Reinmann-Rothmeier 2001], S. 9. Ein Beispiel soll diesen Unterschied verdeutlichen. Quantitative Methoden ermöglichen es etwa, den Grad der Anstrengung, der Freude usw. eines Probanden bzgl. eines Sachverhalts oder einer Situation zu erfragen und so die Effektstärke der Ausprägung eines zuvor festgelegten Merkmals zu erheben. Mit qualitativen Verfahren ist es potenziell möglich, Zugang zu subjektiven, situativen Begründungsmustern zu erschließen. Die Anwendung ergebnisoffener Erhebungsmittel – z.B. offene Fragestellungen in Interviews bzw. Fragebögen – gestatten es dem Probanden, z. B. persönliche Assoziationen oder Gefühle auszudrücken. Das Erfassen derartiger subjektiver Intentionen und Begründungsmuster ermöglicht es wiederum, die Ursachen der Anstrengung, der Freude wesentlich spezifischer herauszuarbeiten [Hey 2001].

Damit wird deutlich, dass bei der Evaluation des netzwerkbasierten Wissensmanagements im Sinne eines integrierenden Forschungsdesigns sowohl qualitative als auch quantitative Methoden eingesetzt und miteinander kombiniert werden sollten, um die Stärken beider Ansätze – die „Objektivität" im Sinne methodischer Standards deduktiver quantitativer Ansätze und die Möglichkeit vertiefter induktiver inhaltlicher explorativer Analysen qualitativer Methoden – zu nutzen. Dies geht konform mit der in der Literatur in zunehmenden Maße aufzufindenden Empfehlung, quantitative Methoden mit qualitativen Methoden zu ergänzen bzw. adäquat miteinander zu kombinieren [Pfister 2004], [Linder 2004], [Schnurer 2005], [Dyson & Campello 2003]. Nur so sei zu gewährleisten, dass Forschungsdesigns der Komplexität des computervermittelten kooperativen Lernens gerecht werden.

Es ist also empfehlenswert, den Untersuchungsgegenstand durch die Anwendung verschiedener Methoden und Indikatoren aus möglichst vielen Perspektiven zu beleuchten [Linder 2004], S. 335. Eine derartige Triangulation, d. h. komplementäre Ergänzung unterschiedlicher Evaluationsmethoden, erleichtert es, verlässliche Schlüsse zu ziehen [Pfister 2004], S. 13.

5 Untersuchungsinstrumente

Nach [Reinmann-Rothmeier & Mandl 2001], S. 133 sind Befragungen und Beobachtungen die meist genutzten Untersuchungsinstrumente zur Evaluation virtueller Lernformen. Im Vergleich zu Face-to-Face-Lernszenarien sind die kooperativen Wissensgenerierungsprozesse in Form von Diskussionsbeiträgen und erarbeiteten Wissensartefakten zudem bereits materiell in der jeweiligen Lernumgebung vorhanden und können z. B. mittels Dokument- und Diskursanalysen untersucht werden. Zusätzlich zu Befragungen und Beobachtungen besteht die Option, das Ergebnis des kooperativen Lernprozesses, den individuellen und kooperativen Lernerfolg, durch Tests oder Dokumentanalysen zu bewerten [Reinmann-Rothmeier & Mandl 2001], S. 133.

Um die verschiedenen Evaluationsinstrumente besser zuordnen und voneinander abgrenzen zu können, können diese in Instrumente zur Ermittlung der Einschätzung der Teilnehmer (Befragungsinstrumente), Instrumente zur Analyse der Lernprozesse (Beobachtungsinstrumente) und Instrumente zur Bewertung der Lernergebnisse (Ergebnisbewertung) kategorisiert werden.

5.1 Instrumente zur Ermittlung der Einschätzung der Teilnehmer

Zentrale Instrumente zur Ermittlung der Einschätzung der Teilnehmer sind: Schriftliche Befragungen (Fragebögen), mündliche Befragungen (Interviews), anekdotische Rückmeldungen. Dabei ist der Einsatz von Instrumenten zur Ermittlung der Einschätzung der Teilnehmer bei der Entwicklung neuer kooperativer Lernszenarien von zentraler Evidenz, denn der Erfolg neuer (kooperativer) Lernarrangements ist ganz entscheidend von der Akzeptanz der Lernenden und deren inhaltlicher Einstufung abhängig [Goertz & Johannig 2004]. Auch wenn die Aussagekraft von Befragungen hinsichtlich Faktoren bzw. Wirkungseffekten wie etwa dem Lernerfolg zweifelhaft bleibt, sind sie doch evident, um die Akzeptanz der Lernenden zu ermitteln. Schriftliche Befragungen bilden dabei das standardisierteste und deshalb am ehesten zu objektivierende Erhebungsinstrument, dessen Ergebnisse,

zumindest hinsichtlich der geschlossenen Items, auch für summative Kausalitäts- bzw. Korrelationsberechnungen verwendet werden können. Freie, unstrukturierte anekdotische Befragungsinstrumente weisen hingegen kaum einen wissenschaftlichen Erkenntnisgewinn auf, können aber im Sinne der formativen Evaluation genutzt werden, um Lernprozess begleitend Hinweise zur fortlaufenden Verbesserung zu gewinnen.

5.2 Instrumente zur Analyse der Lernprozesse

Instrumente zur Analyse der Lernprozesse sind gerade im kooperativen E-Learning von zentraler Bedeutung. Demzufolge ist es nicht verwunderlich, dass in der Literatur ihr Einsatz in der Evaluation stark empfohlen wird, u. a. [Baumgartner 1999]. Ziel der Analyse der Struktur bzw. des Ablaufs der Lernprozesse ist es, den Lernprozess selbst zu verstehen bzw. strukturelle Ausprägungen, Verhaltens- und Handlungsmuster, Regelmäßigkeiten, Problemfelder im Ablauf nachzuvollziehen und zu überprüfen. Im Vergleich zum Face-to-Face-Lernen ist in computerunterstützten Lernszenarien, gerade in asynchronen Foren, die Beobachtung des Ablaufs der Lernprozesse stark durch die Permanenz der Kommunikationsobjekte und Wissensartefakte erleichtert. Zudem stehen in der Regel mit den in der Lernumgebung festgehaltenen Systemnutzungsprotokollen – sogenannte Logfiles, die den Aufruf von Systemfunktionalitäten aufzeichnen – Instrumente zur Verfügung, die das Benutzerverhalten festhalten [Döring 2003], S. 219. Diese Instrumente können als automatische Beobachtungsverfahren aufgefasst werden [Döring 2003], S. 223. Auf dieser technologischen Basis kann versucht werden, das Interaktionsverhalten der Teilnehmer, d. h. sowohl die aktive Kooperation im Diskurs als auch das eher passive rezeptive Nutzungsen zu überprüfen. Zusätzlich zu diesen nicht-reaktiven Verfahren kann das Lerngeschehen ergänzend oder auch alternativ durch verdeckte oder offene teilnehmende menschliche Beobachtungen analysiert werden [Schnell et al. 1999], S. 358.

Instrumente zur Analyse der Kooperation lassen sich in diskursstatische und inhaltsanalytische Diskursbewertungsverfahren differenzieren. Diskursstatistische Ansätze verwenden beispielsweise die Anzahl erstellter Nachrichten pro Zeiteinheit als Indikator für die Aktivität oder die Gliederung der Diskussion als Hinweis zur Bestimmung der inhaltlichen Tiefe eines Diskurses [Stahl & Carell 2004]. [Stahl & Carell 2004] halten solche Ansätze für eine vergleichende Analyse hilfreich, aus sich allein heraus allerdings für wenig aussagekräftig.

Ein Beispiel für diskursstatistische Ansätze stellen etwa die von [Kuhlen 1998] entwickelten Kennzahlen der Informations- und Kommunikationsbereitschaft dar. Kuhlen unterscheidet etwa zwischen absolutem und relativem Informationsgrad als Kennzahlen der Informationsbereitschaft. Dabei berechnet sich der absolute Informationsgrad aus der Anzahl der eingehenden Beiträge pro Zeiteinheit, die einen neuen Thread beginnen. Ein Thread ist hierbei die Gesamtheit aller Beiträge desselben Betreffs. Der relative Informationsgrad berechnet sich aus dem Verhältnis des absoluten Informationsgrads zur Anzahl der Teilnehmer. Beide Indikatoren sieht Kuhlen als Maßzahl für die Bereitschaft zur Wissensteilung. Diskursstatistische Ansätze eignen sich also dazu, ergänzend etwa zu Befragungen oder zur Bewertung der Lernergebnisse, quantifizierbare strukturelle Ausprägungen des Lernprozesses zu messen und auf dieser Basis die Erfüllung/Nichterfüllung grundlegender Qualitätsaspekte wie z. B. den Grad der Beteiligung zu erfassen. Jenseits solcher

grundlegender Aussagen ist aber eine Bewertung der Qualität des Lernprozesse nicht möglich, da statistische bzw. strukturquantitative Eigenschaften aus sich selbst heraus sich nicht notwendigerweise mit Qualität verbinden lassen [Leung 2005].

Inhaltsanalytische Diskursbewertungsverfahren versuchen inhaltlich begründete Hinweise zur Ausprägung von Diskursprozessen zu erschließen und daraus bzw. darauf aufbauend die Qualität der Lernprozesse einzuschätzen. Um eine inhaltlich begründete Einstufung des Diskurses vornehmen zu können, werden Kodierschemata zur Erhebung der problemrelevanten Dimensionen verwendet bzw. entwickelt [Schnell et al. 1999], S. 376. Aufbauend auf dem jeweiligen Kodierschema werden Diskurselemente, beispielsweise Beitragssequenzen, Beiträge und/oder Subelemente von Beiträgen, kategorisiert [Schnurer 2005], S. 4, anschließend quantitativ analysiert und/oder qualitativ interpretiert. Die Aussagekraft der Diskursanalyse ist dabei primär von zwei Faktoren abhängig. Erstens der Passung des verwendeten Kodierschemas, also der Zuverlässigkeit der Aussagekraft der verwendeten Kennzeichnungskategorien hinsichtlich des Erkenntnisinteresses [Archer et al. 2001], zweitens der Stabilität, Wiederholbarkeit und Validität der Zuweisungen der Kennzeichnungskategorien zu den Analysegegenständen [Schnell et al. 1999], S. 376. Hinsichtlich des Kodierschemas ist darauf hinzuweisen, dass im CSCL derzeit keine „Standardkodierschemata“ existieren bzw. verwendete angewandte Kodierschemata kaum in anderen Untersuchungen weitergenutzt werden [Archer et al. 2001]. Bezüglich der Stabilität, Wiederholbarkeit und Validität wird empfohlen, insbesondere die Intersubjektivität bzw. Reliabilität von Zuweisungen dadurch sicherzustellen, dass stets mehrere Kodierer zur Kategorisierung herangezogen und ihre Indexierungskonsistenz geprüft werden soll. Schwierigkeiten, eine hinreichende Interraterkonsistenz zu erreichen, führen bei der Analyse von Diskursen in elektronischen Foren zunehmend dazu, dass die Diskursteilnehmer selbst als Rater fungieren bzw. genutzt werden [Archer et al. 2001].

Beispiele für Kodierschemata existieren im kooperativen E-Learning zuhauf, so nutzen etwa bereits [Baker & Lund 1997] eine Klassifikation in aufgabenbezogene und aufgabenirrelevante Beiträge. U.a. [Schnurer 2005], S. 94-95 differenziert weitergehend zwischen off-task und aufgabenbezogenen Analyseeinheiten. Weitergehend werden letztere in koordinative und inhaltliche Aktivitäten unterteilt. Darauf aufsetzend werden die inhaltlichen Aktivitäten weiter spezifiziert. So wird zusätzlich, durch das Erfassen der Zahl der genannten unterschiedlichen theoretischen Konzepte, das Ausmaß des Einbringens verteilten Wissens operationalisiert. Über die Zuweisung konfliktorientierter bzw. konsensorientierter Kategorien wird schließlich versucht festzustellen, „wie kritisch die Gruppen über die Inhalte diskutieren“ [Schnurer 2005], S. 96. Durch derartige Differenzierung ist es, z. B. in Kombination mit diskursstatistischen Verfahren, möglich herauszufinden, ob beispielsweise ein höherer Anteil inhaltlicher aufgabenbezogener Aktivitäten mit einem höheren Lernerfolg korrespondiert, bzw. zu analysieren, ob sich im Zeitablauf – mit zunehmender Vertrautheit mit der virtuellen Kommunikation – das Verhältnis zwischen koordinativen und inhaltlichen Aktivitäten ändert oder in zunehmendem Maße mehr verteiltes Wissen eingebracht bzw. kritischer diskutiert wird [Schnurer 2005].

Inhaltsanalytische Bewertungsverfahren zielen letztlich dahin, auf semantischer Ebene Indikatoren für die (kognitive) Qualität zu messen. Derartige Erhebungsinstrumente sind komplex. Die Vielzahl unterschiedlicher Untersuchungsdesigns ist kaum erstaunlich, wenn man sich die verschiedenen möglichen evaluativen Blickwinkel auf den kooperativen

Prozess vor Augen führt [Frey et al. 2006]. Methodisch sind derartige Ansätze sehr aufwändig und wenn auch gerade von einer semantischen bzw. kognitiv qualifizierenden Analyse mit Hilfe deduktiv angewendeter bzw. induktiv erarbeiteten Kategorisierungsschemata vertiefte Einblicke in die kooperativen Prozesse erhofft werden, so weist doch [Meyer 2004], S. 113 darauf hin, dass semantische Kodierungsschemata u.U. den Blickwinkel der Untersuchung auch zu verengen vermögen. *„This might argue for regular use of a variety of frameworks, in order to keep the analyst and analysis free from mistaking the world for the lens. This might also prevent one frame becoming the only appropriate form of analysis, avoiding Maslow's caution that "To the man who only has a hammer in the toolkit, every problem looks like a nail."* Des Weiteren ist darauf hinzuweisen, dass aufgrund des hohen zeitlichen Kodieraufwands etliche Studien in diesem Bereich, z. B. [Frey et al. 2006], auf Kontrollmechanismen wie eine Überprüfung der Kodierkonsistenz verzichten bzw. nur einen Kodierer aufweisen, z. B. [Meyer 2004], viele Studien also auch methodisch eher explorativen Charakter besitzen.

Logfile-Analysen werden im kooperativen E-Learning meist supplementär zur Ergänzung bzw. Unterstützung anderer Erhebungsinstrumente, z. B. der Analyse der Kooperation, genutzt. Durch dieses Instrument lassen sich objektive Daten bzgl. Häufigkeit, Zeitpunkt, Dauer von Anwendersitzungen, die Anzahl der Besuche und die Zahl der lesender Zugriffe im System bzw. die Häufigkeit von Funktionsaufrufen ermitteln. Logfile-Analysen stellen damit quasi ein ideales Instrument zur Analyse des Nutzungsverhaltens dar. [Pape et al. 2005] verwenden Logfile-Analysen etwa dazu, um Nutzertypen – Viel- und Wenignutzer – zu differenzieren, Nutzungsmuster und Regelmäßigkeiten zu erschließen und insbesondere auch Nutzungsschwerpunkte und Anlässe zu identifizieren. So ist es möglich, etwa ergänzend zur Analyse des Diskursverhaltens, das Ausmaß und die Kontinuität der nicht aktiv beitragenden Nutzung – Lurking – annähernd nachvollziehen. So kann etwa gemessen werden, wie oft ein Beitrag aufgerufen wurde.

Während Text- respektive Diskurs- und Logfileanalysen auf einer automatischen Datensammlung aufbauen, ermöglichen es verdeckte oder offene teilnehmende Beobachtungen, den Lernprozess aus einer ganzheitlicheren Perspektive [Döring 2003], S. 223 zu betrachten und auch Sachverhalte zu erheben, die nicht explizit in textueller Form automatisch erfasst werden. Ist der Evaluand, der Forscher zugleich Lehrender, so ist er direkt am Lerngeschehen beteiligt und erfasst vor allem im Rahmen der tutoriellen Betreuung technische und didaktische Aspekte des Lernprozesses direkt im Ablauf des Geschehens. Weiterhin kann insbesondere der technologische und zeitliche Aufwand des jeweiligen Lernszenarios festgehalten werden. Generell gilt, dass derartige teilnehmende Beobachtungsverfahren weniger für quantitative Hypothesenprüfung als vielmehr zur Exploration von Problembereichen oder weiteren Forschungsfragen genutzt werden können. Nach [Hey 2001], S. 145 ist die teilnehmende Beobachtung ein Instrument, welches es ermöglicht, vor allem auch festzuhalten, wie etwas geschieht und warum Handlungen entstehen. Der Beobachter steht dabei vor dem Dilemma, dass er zugleich wissenschaftliche Standards beachten und sozial und kulturell verträglich handeln muss.

5.3 Instrumente zur Bewertung der Lernergebnisse

[Reinmann-Rothmeier et al. 2001], S. 134 führt Tests- und Dokumentanalysen als Erhebungsmethoden zur Messung des Output kooperativer Lernszenarien an. Instrumente zur Erhebung des Lernerfolgs beruhen also darauf, dass der Erfolg des Lernens durch prüffähige „Produkte" gemessen wird. Während zur Messung des individuellen Lernerfolgs i. d. R. speziell zu diesem Zweck entworfene Tests bzw. Prüfungsverfahren genutzt werden, wird der kooperative Lernerfolg oft dadurch evaluiert, dass mit Hilfe von Dokumentanalysen das Ergebnisdokument des Kooperationsprozesses geprüft wird [Schnurer 2005], S. 49. Obwohl sich beide Lernergebnisse wechselseitig beeinflussen bzw. voneinander abhängig sind und kooperativer und individueller Lernerfolg konzeptuell kaum zu trennen sind, werden sie bei der Evaluation kooperativer Lernszenarien doch zumeist getrennt erfasst, da bislang keine integrierten Messverfahren existieren und häufig zur Prüfung des Lernerfolgs nur eine der beide Ebenen des Lernerfolgs mit Hilfe von Instrumenten zur Bewertung des Lernergebnisses evaluiert wird [Schnurer 2005].

In Anlehnung an [Schwarz 2001] lässt sich konstatieren, dass der individuelle Lernerfolg kaum objektiv gemessen werden kann und Wissenstests i. d. R. viel zu kurz greifen, da sie sich darauf beschränken zu prüfen, inwieweit Inhalte wiedergegeben werden können. Insbesondere [Schnurer 2005] diskutiert Messverfahren, die zwar auch auf der Analyse von (Test)Ergebnissen beruhen, aber versuchen, elaboriertere Bewertungsverfahren zur Erfassung der Veränderung individueller kognitiver Strukturen einzusetzen. Diese Verfahren basieren dabei i. d. R. darauf, dass zunächst eine taxonomische Differenzierung des Lernerfolgs vorgenommen wird. Ein Beispiel einer solchen Taxonomie stellt etwa die genannte Lernzielkategorisierung von [Bloom 1972][3] dar, die auf einer sechsstufigen hierarchisch aufeinander aufbauenden Skala die Komplexität von Lernzielen differenziert. Die einzelnen Kategorien lassen sich nutzen, um gezielt verschiedene qualitative Ebenen des Lernerfolgs analytisch zu fassen und bei der Evaluation zu prüfen. Im Grunde handelt es sich also bei dieser Art der Ergebnisbewertung um nichts Anderes als eine spezielle Anwendung der oben dargestellten Inhaltsanalyseverfahren.

Die Evaluation des Lernerfolgs anhand des Erfüllungsgrades sprachlich formulierbarer summativer Bewertungskriterien ist nicht nur hinsichtlich der wissenschaftlichen Evaluation von Lernszenarien State of the Art, sondern bildet das Rückgrat des gesamten Bildungssystems nicht nur in Deutschland. In Feldszenarien des kooperativen Lernens lässt sich aus pragmatischer Sicht deshalb auch die reale Leistungsbewertung, die durch die Lehrenden vorgenommen wird, als Grundlage der wissenschaftlichen Bewertung des Lernergebnisses verwenden – vgl. hierzu auch das Untersuchungsdesign von [Holl 2003]. Wird die Leistungsbewertung nur von einer Person wahrgenommen, stellt sich aus evaluationsmethodischer Perspektive die Frage der Intersubjektivität der Leistungsbewertung. Ande-

[3] A) Wissen: Reproduktion von Fakten.
B) Verstehen: Überblick über Ereignisse, Informationen. Ableitung von Implikationen und Konsequenzen.
C) Anwenden: Übertragung von Sachverhalten in andere Zusammenhänge.
D) Analyse: Erkenntnis der Struktur von Sachverhalten. Konstruktion von Zusammenhängen zwischen Konzepten.
E) Synthese: Verknüpfung, Zusammenfassung von inhaltlich zusammenhängenden Aussagen, Aufbau neuer Wissensstrukturen.
F) Evaluation: Bewertung auch komplexer Zusammenhänge und Strukturen.

rerseits kann bei der Leistungsbewertung durch einen Lehrenden zugleich eine hohe Validität der Leistungsanalyse erwartet werden, da Lehrende i. d. R. per Definition als kompetente Experten sowohl der Beurteilung inhaltlicher als auch prozeduraler Lernziele einzustufen sind. U.a. [Lind 2004] weist darauf hin, dass zur Prüfung des Lernzuwachses insbesondere eine Vorher-Nachher-Messung vorgenommen werden soll. Da in realen Kursszenarien der Einsatz von Kontrollgruppen kaum möglich ist bzw. kaum empfohlen werden kann, ist aber auch bei einer Vorher-Nachher-Messung nicht auszuschließen, dass der gemessene Lernerfolg durch andere, außerhalb des Lernszenarios liegende Ursachen, etwa informelles Lernen [Overwien 2004], begründet ist. Im Unterschied zur Erhebung des individuellen Lernerfolgs anhand von wie auch immer gearteten Tests ist die Analyse des kooperativen Lernerfolgs weitgehend auf die Inhaltsanalyse der virtuellen Diskursobjekte bzw. Wissensartefakte beschränkt, die als Ergebnis des Lernprozesses betrachtet werden.

6 Ergebnis: Reichweite und Grenzen der Evaluierbarkeit kooperativer E-Learning-Szenarien

In den vorhergehenden Kapiteln wurden Zweck und Ziel von Evaluationen im kooperativen E-Learning geschildert, methodische Aspekte diskutiert und Untersuchungsinstrumente zur Erhebung evaluationsrelevanter Daten und Zusammenhänge dargestellt. Damit wurden konkrete Möglichkeiten der Triangulation verschiedener Werkzeuge im Sinne einer Feedback-Prozess-Produkt-Analyse erschlossen. Zusammenfassend bleibt anzumerken, dass es sinnvoll ist, bei der Evaluation des netzwerkbasierten Wissensmanagements in Hochschulkursen sowohl qualitative als auch quantitative Instrumente zur Erhebung zu nutzen und dabei möglichst alle Instrumente aus den genannten Bereichen zu kombinieren und miteinander in Bezug zu setzen, um aussagekräftige Ergebnisse zu erzielen.

Es ist wichtig zu verdeutlichen, dass die Evaluation des netzwerkbasierten Wissensmanagements in der Hochschullehre nicht nur primär darauf zielt, ein erprobtes methodisches Instrumentarium mehr oder weniger ressourcenintensiv anzuwenden, sondern dass die Evaluation bzw. die Instrumente sowohl in ihren Ausprägungen – beispielsweise die Ausgestaltung der Diskursbewertungsverfahren – als auch in ihrer Kombination (Triangulation) selbst als Forschungsgegenstand zu begreifen ist. Untersuchungsdesigns sind deshalb nicht nur hinsichtlich der Gültigkeit und Zuverlässigkeit der Ergebnisse zu hinterfragen. Vielmehr ist es auch sinnvoll, die eingesetzten Methoden und ihre Kombination hinsichtlich ihrer Zuverlässig (Validität) in Bezug zu den Evaluationszielen kritisch zu reflektieren. Dabei ist a priori davon auszugehen, dass Evaluationsergebnisse explorativer Natur sind. Zumal nach Schulmeister insbesondere nicht nur die Technologie und Didaktik, sondern die Lehrenden selbst einen wesentlichen Wirkungsfaktor in Lernszenarien darstellen [Schulmeister 2002], S. 402, der in Evaluationen berücksichtigt werden muss. Schließlich ist darauf hinzuweisen, dass auch bei dem denkbar sorgfältigsten Evaluationsdesign kursexterne Gegebenheiten nicht erfasst und somit wesentliche Wirkungsfaktoren auf den Lernerfolg im Kurs, etwa die zeitliche Belastung durch andere Kurse oder sonstige Merkmale der Individualsphäre der Lernenden, nicht berücksichtigt werden [Kromrey 2001]. Schlussendlich bleibt anzumerken, dass Lernszenarien im Sinne von Lernangeboten nur geeignet sind, Lernprozesse zu erleichtern, und diese nicht kausal bewirken, also erzwingen können [Kromrey 2001].

Literaturverzeichnis

Archer, W.; Garrison, D. R.; Anderson, T.; Rourke, L. (2001). A framework for analysing critical thinking in computer conferences, http://www.ll.unimaas.nl/euro-cscl/Papers/6.doc (letzter Zugriff 23.08.2006).

Baker, M. J.; Lund, K. (1997). Promoting reflective interactions in a computer-supported collaborative learning environment. Journal of Computer Assisted Learning, 13 Nr. 175, 193.

Baumgartner, P. (1999). Evaluation vernetzten Lernens: 4 Thesen, http://paedpsych.jk.uni-linz.ac.at/PAEDPSYCH/EVALUATION/EVALUATIONLITORD/Baumgartner99.pdf (letzter Zugriff 14.02.2006).

Bloom, B. S. (1972). Taxonomie von Lernzielen im kognitiven Bereich. Weinheim: Beltz.

Döring, N. (2003). Sozialpsychologie des Internet. Die Bedeutung des Internet für Kommunikationsprozesse, Identitäten, soziale Beziehungen und Gruppen. Göttingen: Hogrefe Verlag für Psychologie.

Dyson, M. C.; Campello, S. B. (2003). Evaluating Virtual Learning Environments: what are we measuring? Electronic Journal of e-Learning, 1 Nr. 1, 11-20.

Frey, B. A.; Sass, M. S.; Alman, S. W. (2006). Mapping MLIS Asynchronous Discussions. International Journal of Instructional Technology & Distance Learning, 3 Nr. 1, 3-16, http://www.itdl.org/Journal/jan_06/article01.htm (letzter Zugriff 25.02.2006).

Friedrich, H. F; Hesse, F. W. (2001). Partizipation und Interaktion im virtuellen Seminar - ein Vorwort. In: Partizipation und Interaktion im virtuellen Seminar. Friedrich, H. F.; Hesse, F. W. (eds.). Münster, New York, München, Berlin: Waxmann, 7-11.

Goertz, L.; Johannig, A. (2004). Das Kunststück, alle unter einen Hut zu bringen. Zielkonflikte bei der Akzeptanz des E-Learning. In: Was macht E-Learning erfolgreich? Grundlagen und Instrumente der Qualitätsbeurteilung. Tergan, S. –O.; Schenkel, P. (eds.). Berlin, Heidelberg, New York, Hongkong, London, Mailand, Paris, Tokio, Wien: Springer-Verlag, 83-92.

Griesbaum, J. (2006). Joachim Griesbaum: Mehrwerte des Kollaborativen Wissensmanagements in der Hochschullehre – Integration asynchroner netzwerkbasierter Szenarien des CSCL in der Ausbildung der Informationswissenschaft im Rahmen des K3-Projekts. Dissertation Fachbereich Informatik und Informationswissenschaft, Universität Konstanz, Konstanz.

Griesbaum, J. (2004). Curriculare Vermittlung von Informationskompetenz: Konzepte, Ziele, Erfahrungen eines experimentellen Retrievalkurses (K3). Konstanz: UVK, 283-299.

Griesbaum, J.; Rittberger, M. (2005). A Collaborative Lecture in Information Retrieval for Students at Universities in Germany and Switzerland. In: Proceedings of the World Library and Informa-tion Congress: 71st IFLA General Conference and Council. "Libraries - A voyage of discov-ery", http://www.ifla.org/IV/ifla71/papers/068e-Griesbaum_Ritterberg.pdf (letzter Zugriff 26.08.2006)

Haake, J. M.; Schwabe, G.; Wessner, M. (2004). Entwicklungsprozess. In: CSCL-Kompendium. Lehr- und Handbuch zum computerunterstützten kooperativen Lernen. Haake, J. M.; Schwabe, G.; Wessner, M. (eds.). Oldenbourg: Springer, 288-294.

Hey, A. H. (2001). Feedback und Beurteilung bei selbstregulierter Gruppenarbeit. Berlin: Dissertation.de.

Holl, B. (2003). Entwicklung und Evaluation eines Unterrichtskonzeptes für computergestütztes kooperatives Lernen. Computer Supported Cooperative Learning (CSCL) am beruflichen Gymnasium für Informations- und Kommunikationstechnologie, http://archiv.tu-chemnitz.de/pub/2004/0021/data/Dissertation_Holl.pdf (letzter Zugriff 25.08.2006).

Kerres, M. (2001). Multimediale und telemediale Lernumgebungen. München: Oldenbourg Verlag.

Kromrey, H. (2001). Studierendenbefragungen als Evaluation der Lehre? Anforderungen an Methodik und Design. In: Hochschulranking. Zur Qualitätsbewertung von Studium und Lehre. Engel, U. (ed.). Frankfurt a. M.; New York: Campus Verlag, 11-47.

Kuhlen, R. (2006). In Richtung Summarizing für Diskurse: In: Information und Sprache. Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern Festschrift für Harald H. Zimmermann. Herausgegeben von Ilse Harms, Heinz-Dirk Luckhardt und Hans W. Giessen K·G·Saur München, S. 55-74.

Kuhlen, R. (2002). Vorhabensbeschreibung K3 - Wissensmanagement über kooperative verteilte Formen der Produktion und der Aneignung von Wissen zur Bildung von konzeptueller Informationskompetenz durch

Nutzung heterogener Informationsressourcen, http://www.k3forum.net/vorhabensbeschreibung.pdf (letzter Zugriff 28.08.2006).

Kuhlen, R.; Griesbaum, J.; Jiang, T.; König, J.; Lenich, A.; Meier, P.; Schütz, T.; Semar, W. (2005). K3 - an e-Learning Forum with Elaborated Discourse Functions for Collaborative Knowledge Management. In: Proceedings of E-Learn 2005 World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education October 24-28, Vancouver BC, Canada. 2981-2988.

Leung, K. H. (2005). A critical review of current research on on-online collaborative problem-based learning. Proceedings of E-Learn 2005. World Conference on E-Learning in Corporate, Government, Healthcare, Higher Education, Oct 24-28 2005 Vancouver BC, Canada: 1683-1690.

Linder, U. (2004). Qualitätssicherung. In: CSCL-Kompendium. Lehr- und Handbuch zum computerunterstützten kooperativen Lernen. Haake, J. M.; Schwabe, G.; Wessner, M. (eds.). Oldenbourg: Springer, 326-340.

Mayring, P. (1996). Einführung in die qualitative Sozialforschung. Weinheim: Psychologie Verlags Union.

Meyer, K. A. (2004). Evaluating Online Discussions: Four Different Frames of Analysis. Journal of Asynchronous Learning Networks, 8 Nr. 4, 101-114, http://www.aln.org/publications/jaln/v8n2/pdf/v8n2_meyer.pdf (letzter Zugriff 25.08.2006).

Overwien, B. (2004). Internationale Sichtweisen auf "informelles Lernen" am Übergang zum 21. Jahrhundert. In: Ganztagsbildung in der Wissensgesellschaft. Otto, H.; Coelen, T. (eds.). Wiesbaden: Verlag für Sozialwissenschaften, 51-73.

Pape, B.; Janneck, M.; Klein, M. (2005). Logfile-Analysen zur Evaluation der didaktischen Einbettung von CSCL-Systemen am Beispiel der CommSy-Nutzung in offenen Seminaren. e-learning and education (eleed) Nr. 1, http://eleed.campussource.de/archive/1/85/ (letzter Zugriff 22.02.2006).

Pfister, H. -R. (2004). Forschungsmethoden. In: CSCL-Kompendium. Lehr- und Handbuch zum computerunterstützten kooperativen Lernen. Haake, J. M.; Schwabe, G.; Wessner, M. (eds.). Oldenburg: Springer, 5-13.

Reinmann-Rothmeier, G.; Mandl, H. (2001). Virtuelle Seminare in Hochschule und Ausbildung. Bern: Hans-Huber.

Reinmann-Rothmeier, G. (2001a). Wissensmanagement in der Forschung. Gedanken zu einem integrativen Forschungsszenario. Forschungsbericht Nr. 132.

Reinmann-Rothmeier, G.; Mandl, H.; Nistor, N.; Neubauer, A.; Erlach, C.; Weinberger, A.; Lerche, T. (2001). Evaluation virtueller Seminare in Schule und Hochschule. In: Virtuelle Seminare in Hochschule und Weiterbildung. Drei Beispiele aus der Praxis. Reinmann-Rothmeier, G. & Mandl, H. (eds.). Bern: Hans Huber, 131-150.

Schnell, R.; Hill, P. B.; Esser, E. (1999). Methoden der empirischen Sozialforschung. München: Oldenbourg

Schnurer, K. (2005). Kooperatives Lernen in virtuell-asynchronen Hochschulseminaren. Eine Prozess-Produkt-Analyse des virtuellen Seminars "Einführung in das Wissensmanagement" auf der Basis von Felddaten. Berlin: Logos Verlag.

Schulmeister, R. (2002a). Grundlagen hypermedialer Lernsysteme. Theorie - Didaktik - Design. München: Oldenbourg Verlag.

Schwarz, C. (2001). Evaluation von e-learning in der Hochschullehre. Ein Experimentierfeld im Experimentierfeld. In: Evaluation - Reformmotor oder Reformbremse? Deutsche Gesellschaft für Evaluation (ed.). Köln: DeGEval.

Stahl, G.; Carell, A. (2004). Kommunikationskonzepte für eine CSCL-Didaktik, https://web-imtm.iaw.ruhr-uni-bochum.de/pub/bscw.cgi/d268933/30410.pdf (letzter Zugriff 22.08006).

Weinberger, A. (2003). Scripts for Computer-Supported Collaborative Learning. Effects of social and epistemic cooperation scripts on collaborative knowledge construction. München: LMU München: Fakultät für Psychologie und Pädagogik, http://edoc.ub.uni-muenchen.de/archive/00001120/01/Weinberger_Armin.pdf (letzter Zugriff 28.08.2006).

Wessner, M.; Pfister, H. -R. und Miao, Y. (1999). Using Learning Protocols to Structure Computer-Supported Cooperative Learning. In: Proceedings of ED-MEDIA 99 - World Conference on Educational Multimedia, Hypermedia & Telecommunications. 471-476.

Die Suchmaschine SENTRAX. Grundlagen und Anwendungen dieser Neuentwicklung

Hans-Joachim Bentz

Universität Hildesheim
Institut für Mathematik und Angewandte Informatik
Marienburger Platz 22
31141 Hildesheim
bentz@cs.uni-hildesheim.de

Zusammenfassung

Es wird eine intelligente Suchmaschine für den bequemen Zugriff auf strukturierte und unstrukturierte Informationen vorgestellt. Grundlage bilden 4 verschiedene Ähnlichkeitsmaße auf den Datensorten in der Datenbasis gemäß den jeweiligen Aufgaben: Schreibweisentolerante Suche, Kontextähnliche Suche, Zugriff auf Dokumententreffer, Doublettensuche.

Abstract:

This article introduces an automatic "essence-extractor-engine" which works both on structured and inhomogenous document collections and supports interactive searching.

1 Einleitung

An einem vernetzten Lern- bzw. Arbeitsplatz, wo sowohl unterschiedliche Aufgaben, Anforderungen und Prozesse wirken als auch verschiedenartige Arbeitsmaterialien und Datenbasen in den Zugriff gestellt werden, ist die "ad hoc-Suche" nur selten erfolgreich. Unter anderem liegt es an den Schwächen herkömmlicher Suchtechnik, welche meist auf Matching-Algorithmen auf sogenannten "invertierten Listen" beruht. Diese Listen enthalten praktisch alle Wörter aus den Texten, wegen des schnellen Zugriffs alphabetisch sortiert, zusammen mit den Herkunftskoordinaten. Weicht der Suchbegriff auch nur wenig vom Listenwort ab, insbesondere im ersten Wortteil, wie zum Beispiel Apartment und Appartement, dann geht die Suche fehl oder bleibt unvollständig. Man hat mehrere Abhilfen versucht (wie etwa eine Rückwärtstrunkierung, wohinter sich eine invertierte Liste aller Wörter von hinten gelesen verbirgt; oder eine Reduktion der Wörter auf ihren Wortstamm samt Zerlegung in Wortbestandteile, wie Arbeit-s-Material-ien; oder auch zusätzliche Verweise etc.), jedoch greift keine davon entscheidend durch. Die SENTRAX Engine verfolgt einen anderen Ansatz: Texte werden als abstrakte Sequenzen von Strings über einem festzulegenden Alphabet interpretiert und darauf dann Mustererkennung betrieben. Je nach Aufgabenstellung setzt eine passende Routine ein, die mit einem auf die Problemstellung zugeschnittenen Ähnlichkeitsmaß Muster sucht und erkennt. Somit

werden (auf Stringebene) Tippfehler bzw. Schreibvarianten toleriert oder (auf Wort- bzw. Satzebene) bedeutungsverwandte Begriffe zugelassen. Es werden vier Aufgabenstellungen unterschieden und unterstützt:

1) *lexico*: die Suche nach ähnlichen Zeichenketten,
2) *context*: die Suche nach bedeutungsverwandten bzw. im gleichen Kontext vorkommenden Begriffen,
3) *treffer*: das auf Wortebene tolerante Holen von Dokumenten mit den Suchbegriffen,
4) *similardoc*: das auf Dokumentebene tolerante Präsentieren von Doubletten oder Fastdoubletten.

SENTRAX steht für Essence Extractor Engine. Das Hauptmotiv bei der Entwicklung der Suchmaschine war, das Problem der meist unscharfen Konzeptbeschreibung in den Griff zu bekommen. Für den Benutzer ist es immer schwierig, bei seiner Anfrageformulierung geeignete Begriffe zu finden, die einerseits sein (vages) Informationsbedürfnis wiedergeben und andererseits die geeignete Grundlage für die Suche im System bieten. Wegen der Vielzahl an Möglichkeiten, ein und denselben Vorgang oder Sachverhalt zu beschreiben, ist es nicht leicht, mit einer begrenzten Anzahl von Suchbegriffen die Weite eines Konzepts zufriedenstellend abzudecken. Um hierbei zu helfen befindet sich im SENTRAX-Container eine (maschinelle) Zusammenstellung solcher Begriffscluster aus den Texten, die geeignet sind bestimmte Konzepte zu verkörpern. Über eine mehrstufige Kookkurrenzanalyse (basierend auf den Untersuchungen von Wettler et.al [1995] und der Arbeit von Ackermann [2000]) werden „Wortwolken“ gebildet, die die „Essenz der Texte“ repräsentieren sollen. Die mehrdimensionale Wortwolke wird dann zweidimensional auf den Bildschirm projiziert.

2 Typische Hindernisse bei herkömmlicher Technologie

In der nachfolgenden Übersicht sind ein paar Beispiele von Wörtern und Wortkonstruktionen zusammengestellt, die Probleme bei der computergestützten Suche mit sich bringen. Alle Beispiele entstammen realen Dokumenten, sind also in der Praxis vorgekommen. Die Gruppierung ist willkürlich vorgenommen, teilweise überschneiden sich die Typen. Sie sollten sich ohne größere Erläuterung verstehen lassen. Lediglich das Beispiel "method-rnethod" bedarf eines Kommentars: in einer PDF-Datei stand tatsächlich "r-n" anstatt m (OCR Fehlgriff?), weshalb das lesende Auge hier keinen "Fehler" sah, die SENTRAX Engine jedoch einen solchen auswies. Bei genauerer Inspektion löste sich das wie beschrieben auf. Die Phänomene, ob nun tückisch oder simpel, sind natürlich nicht nur auf deutsche Texte beschränkt, vergleichbare Fälle gibt es auch in anderen Sprachen.

Suchworte sind nicht exakt so im Text enthalten, wie erwartet
(Herrmann–Hermann, Ullrich-Ulrich, Detlef–Detlev, Maßstab–Massstab, Notfallmaßnahme–Notfallschutzmaßnahme, Uranbergbau–Uranerzbergbau)

Suchworte haben zulässige oder tolerierte Schreibvarianten
(Foto – Photo, Fahrkostenerstattung – Fahrtkostenerstattung, grey – gray, Potenzial – Potential, Apartment-Appartement, Bundesforschungsminister – Bundesminister für Forschung und Technologie, Numerierung - Nummerierung)

Suchworte sind im Singular, im Dokument aber im Plural oder umgekehrt

(Visum – Visa, Universum - Universen)

Suchworte sind zwar korrekt eingegeben, im Dokument aber fehlerhaft
(Archivierung-Archiverung, Libyen–Lybien, method-rnethod)

Eingabe ist falsch geschrieben (Mitterand) **oder hat Tippfehler**
(Grundwasserstömung, refernce)

Anstelle des Suchworts steht „leider" ein bedeutungsverwandter Begriff im Text
(Fusion–Zusammenschluß-Verschmelzung, Wegbeschreibung–Anfahrtplan, Firmenpleiten–Konkurse–Insolvenzen)

3 Vom Suchen zum Finden durch vier Optionen

In diesem Abschnitt wird kurz die "normale" Arbeitsweise mit den SENTRAX-Funktionen beschrieben. Dazu stelle man sich vor, einen Datenbestand zu haben, der aus recht unstrukturiertem Material bestehen mag, zum Beispiel aus x1000 Office-Dokumenten (HTML, TXT, WORD, PDF, PPT, EXCEL). Man kann sie im File-System liegen haben oder auch in einer Datenbank. Die SENTRAX-Engine geht in der Standardanwendung alle Dokumente einmal durch und "liest" sie zum Zweck der Erstellung eines "Containers", der einem Index entspricht. Aus praktischen Gründen wird von jedem Dokument eine HTML-Version hergestellt. Obgleich dadurch manchmal gewisse Formatierungen und Objekte verloren gehen, hat man zum Ausgleich erstens einen sehr schnellen Zugriff auf den Text - auch über das Inter- oder Intranet- (man braucht also nicht eigens eine Applikation, z.B. ein WORD-Programm, zu öffnen) und zweitens die Möglichkeit die Fundstellen anzuspringen und hervorzuheben (z.B. durch eine Highlight-Funktion). Wer einmal versucht hat, durch ein 10 MByte großes PDF-Dokument zu blättern, weiß die beschriebene Option zu schätzen! Nachdem eine später als Trefferkandidat angesehene HTML-Version positiv geprüft wurde, kann natürlich mit einem Klick das zugehörige Original aufgerufen und ggf. bearbeitet werden (vgl. unten: (5) Ansichtsoptionen).

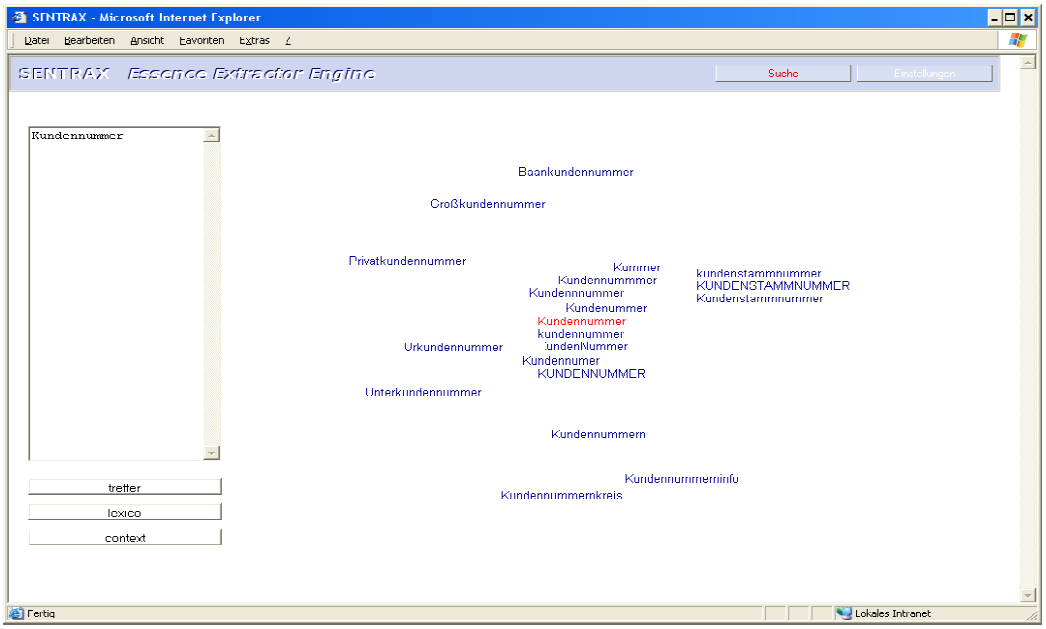

Abb. 1: SENTRAX Bildschirm nach Eingabe: "Kundennummer" und Aktivierung der *lexico*-Funktion

(1) Die *lexico*-Funktion

Bei der ersten Eingabe empfiehlt sich, das Suchwort der *lexico*-Funktion zu übergeben. Man erhält dann eine (manuell einstellbare) Anzahl von Begriffen, die String-ähnlich zur Eingabe sind und alle im Dokumentenbestand vorkommen. Sie werden in einer Map zweidimensional dargestellt, wobei sich die Kandidaten um das ins Zentrum platzierte Suchwort, das farblich hervorgehoben ist, gruppieren. Je näher dran, desto ähnlicher. Diese Ausgabe kann auch (manuell einstellbar) in Form einer Liste gegeben werden. Besonders hilfreich ist die *lexico*-Funktion in dem Fall, wo das Eingabewort nicht in der Datenbasis vorkommt. Man sieht diesen Umstand sofort, hat aber automatisch die Menge der Angebote von schreibweisenähnlichen, aus denen man sich dann eines oder mehrere für die weiteren Suchschritte wählen kann. Überdies erkennt man stets vorkommende Tippfehler, Schreibvarianten oder Begriffsvariationen.

So erkennt man in dem Beispiel in Abbildung 1 erstens: das Eingabewort "Kundennummer" gibt es in den Daten, zweitens: es kommt außerdem noch mit Tippfehler vor (Kundenummer), drittens: es gibt diverse Kompositionen, die für die Suche interessant oder hilfreich sein können (Kundenstammnummer, Privatkundennummer, Kundennummernkreis etc.).

Durch Anklicken eines anderen Worts wird dessen Umgebung auf den Schirm geholt, durch nochmaliges Anklicken (toggeln) eines bereits gefärbten Worts wird dieses wieder ausgeschaltet. Durch zusätzliche Eingabe eines Wortes ins Eingabefenster kommt dessen Umgebung dazu. Nachdem man sich für ein oder mehrere Wörter entschieden hat, die für die weiteren Schritte verwendet werden sollen, kann man entweder die *context*-Funktion oder die *treffer*-Funktion aktivieren.

(2) Die *context*-Funktion

Nach Eingabe von Begriffen oder Auswahl aus der lexico-Map kann die *context*-Funktion aktiviert werden. Sie projiziert eine (im Begriffsumfang manuell einstellbare) Sammlung von Wörtern aus dem Container auf den Bildschirm. Diese Sammlung beinhaltet solche Begriffe, die prädominant im Kontext mit den Suchwörtern der Eingabe vorkommen. Dabei sind auch höhere Ordnungen mitberechnet, weshalb es passieren kann, dass zwei Begriffe als nahe eingestuft und gezeigt sind, obwohl sie nie zusammen in demselben Dokument vorkommen. Auch hier kann man wieder weitere vom Bildschirm dazuklicken oder wegklicken. Damit wird eine interaktive, zielgerichtete Suche möglich. Insbesondere zeigt sich bei den meisten Nutzern, daß die angebotenen Wortgruppen Teilkonzepte oder verbundene Konzepte aufrufen, also wie ein "passives Gedächtnis" fungieren. Somit leistet die SENTRAX Technologie einen wichtigen Beitrag zur vollautomatischen Erschließung von Sinnstrukturen im Datenbestand.

Ein treffendes Beispiel für die diversen Eigenschaften zeigt die Abbildung 2. Der Bildschirm ist einer Arbeit von Kummer [2006] entnommen, wo Trefferlisten herkömmlicher Suchmaschinen mit Hilfe der SENTRAX-Technologie analysiert werden. Im konkreten Beispiel wurde das Wort "Todsünden" in eine der Internetsuchmaschinen eingegeben und ein Satz der angebotenen Trefferseiten dann in einem SENTRAX-Container erfasst. Das gleiche Suchwort lieferte nun (über die *context*-Funktion) den gezeigten

Screen. Man kann mit einem Schlag alle Todsünden erkennen als auch weitere Begriffe, die auf den fraglichen Seiten im Zusammenhang mit dem Suchkonzept stehen.

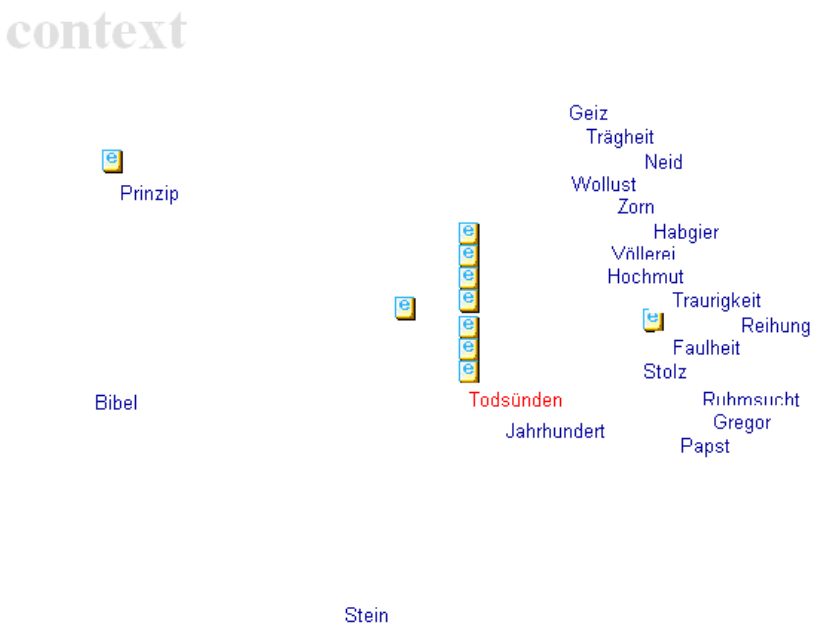

Abb. 2: SENTRAX Bildschirm nach Eingabe: "Todsünden" mit der *context*-Funktion (aus Kummer [2006])

Im Beispiel der Abb. 3 wurden zwei Wörter -"kreative Selbstpräsentation"- in eine der Internetsuchmaschinen eingegeben und wieder ein Anfangsstück der gemeldeten Treffer-

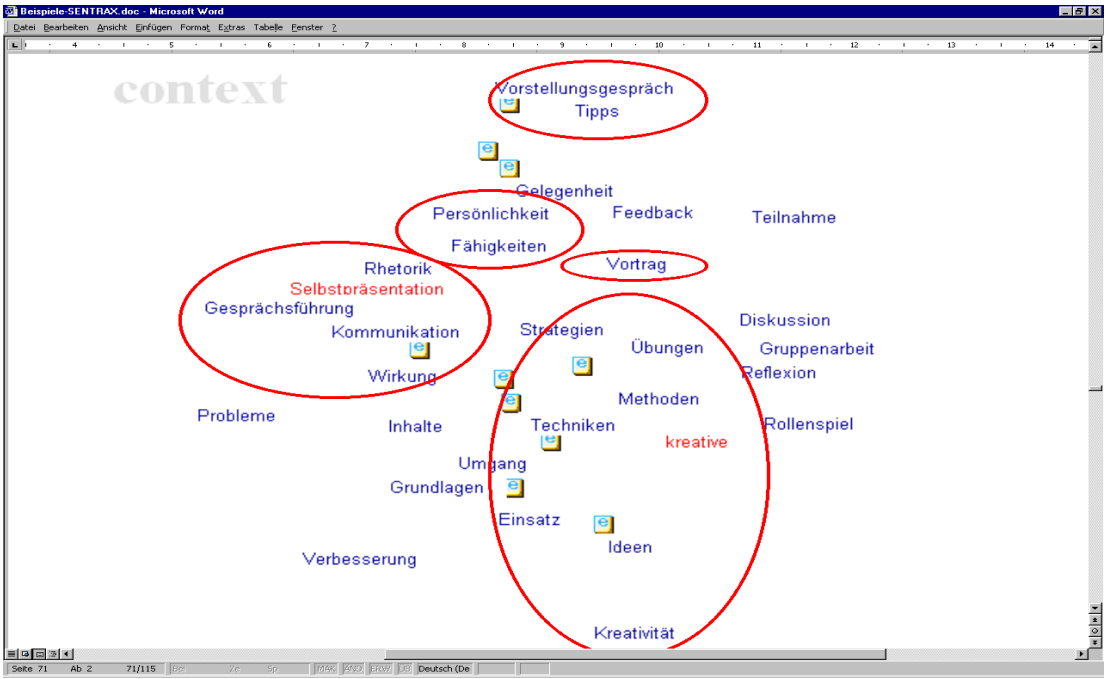

Abb. 3: SENTRAX *context*-Screen nach Eingabe: "kreative Selbstpräsentation" (aus Kummer [2006])

seiten mit der SENTRAX verarbeitet. Man kann ganz gut mehrere Teilcluster erkennen, die unterschiedliche Aspekte der "kreativen Selbstpräsentation" darstellen. Die Markierungen sind hier manuell zum Zweck der besseren Illustration ergänzt.

(3) Die *treffer*-Funktion

Normalerweise gibt es zu einem Suchwort viele Trefferdokumente, in denen es mit dieser oder jener Bedeutung enthalten ist. Ein Teil davon wird durch entsprechende Symbole bereits auf dem *context*-Screen gezeigt und kann von dort direkt aufgerufen werden. Dieses Vorgehen ist im allgemeinen aber nicht besonders effizient. Man will ja wenn möglich vermeiden, manuell viele Dokumente durchzulesen um zu sehen, ob sie als "Lösung" des Suchproblems in Frage kommen. Vielmehr hat man durch die bequeme Komplettierung des gesuchten Konzepts mittels Hinzuklicken passender Wörter aus der *context*-Wortwolke die Möglichkeit, die Anzahl der in Frage kommenden Dokumente einzuschränken. Je spezifischer nämlich die Beschreibung der Suchidee wird, um so weniger Treffer wird es normalerweise geben. Sobald die Anzahl der Dokument-Icons übersichtlich geworden ist, kann man sich mit der treffer-Funktion spezifischere Informationen zeigen lassen und einen Schnellzugriff auf ein eventuell interessierendes Dokument vorbereiten.

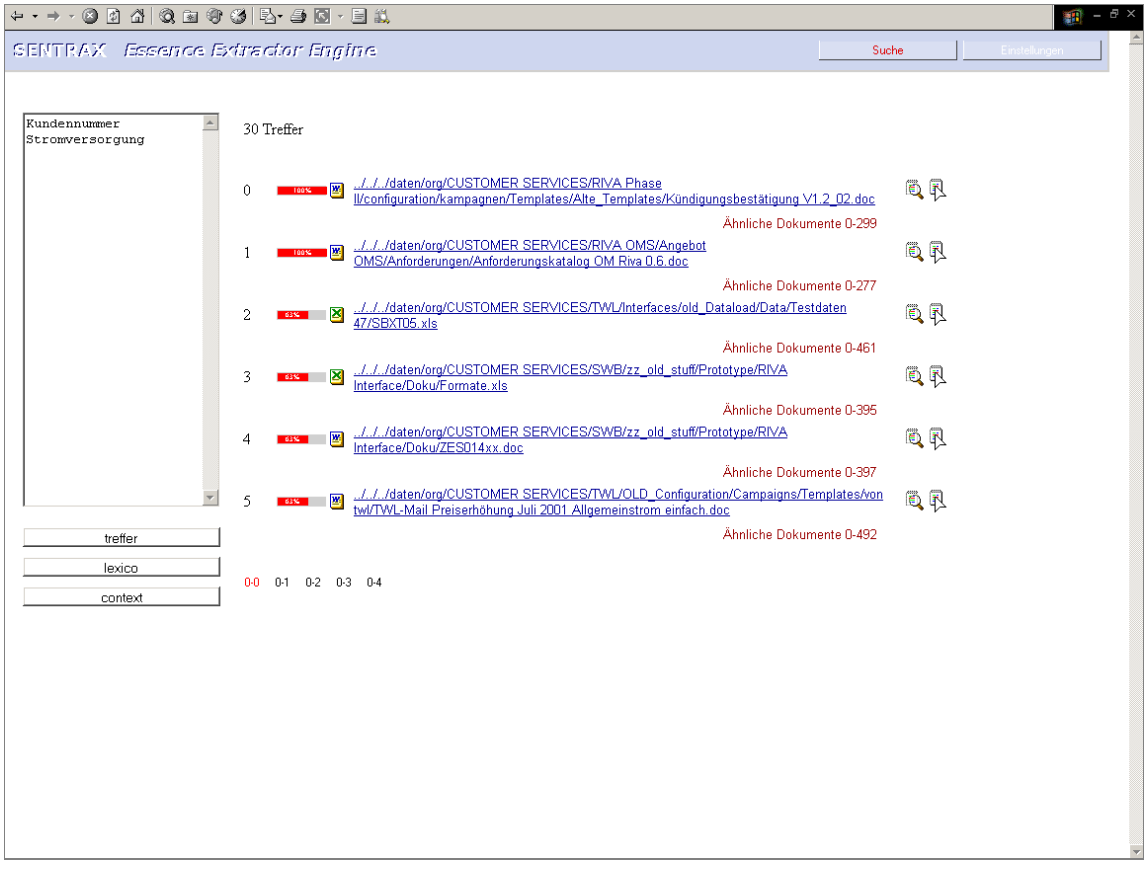

Abb. 4: SENTRAX *treffer*-Screen nach Eingabe: "Kundennummer Stromversorgung"

Im Beispiel der Abbildung 4 sieht man die ersten sechs von 30 gemeldeten Treffern mit Dokumentname und Pfad und weiteren Optionen. Es ist auch einstellbar, die z.B. ersten

300 Zeichen aus dem Text gezeigt zu bekommen und weitere Metainformationen (wie Autor, Dokumentgröße usw.), um sich in diesem Stadium der Suche bequemer orientieren zu können. Dieser Ausgabescreen entspricht in etwa der Situation bei herkömmlicher Internetsuche: Man gibt etwas ein und erhält eine Liste von "Treffern".

In der konkreten Suche der Abbildung 4 gab es zwei Eingabewörter deren Vorkommen im jeweiligen Trefferdokument durch den Füllgrad des roten Balkens angezeigt wird. Ist er voll gefüllt, dann sind alle Eingaben enthalten, ist er nur teilgefüllt, dann fehlt eines usw. Die Ausgabeliste wird standardmäßig nach dem Füllgrad sortiert. Die Symbole rechts in den Zeilen sind Buttons und erlauben die schnelle Ansicht des Dokuments im HTML-Format. Ein Klick auf das Dokument selbst öffnet dieses.

(4) Die Funktion "Ähnliche Dokumente" (*similarDoc*)

Diese Funktion ist die letzte der vier SENTRAX-Suchoptionen. Hier beruht das Ähnlichkeitsmaß auf einer gewichteten Anzahl gemeinsamer (bedeutungsvoller) Wörter. Man aktiviert diese Funktion für ein Dokument, indem man im *treffer*-Screen das einem Dokument zugehörige Feld "Ähnliche Dokumente" anklickt. Es muss nicht das oberste Dokument sein, sondern kann beliebig aus der Trefferliste gewählt werden.

Als Folge wird eine Liste angezeigt, die vom aktivierten Ausgabedokument angeführt wird. Darunter stehen dann mit absteigender Ähnlichkeit die Kandidaten. Der Grad der Ähnlichkeit, den die Engine ermittelt, wird wieder durch die Füllung des Balkens bzw. auch durch den darin eingetragenen %-Wert ausgedrückt.

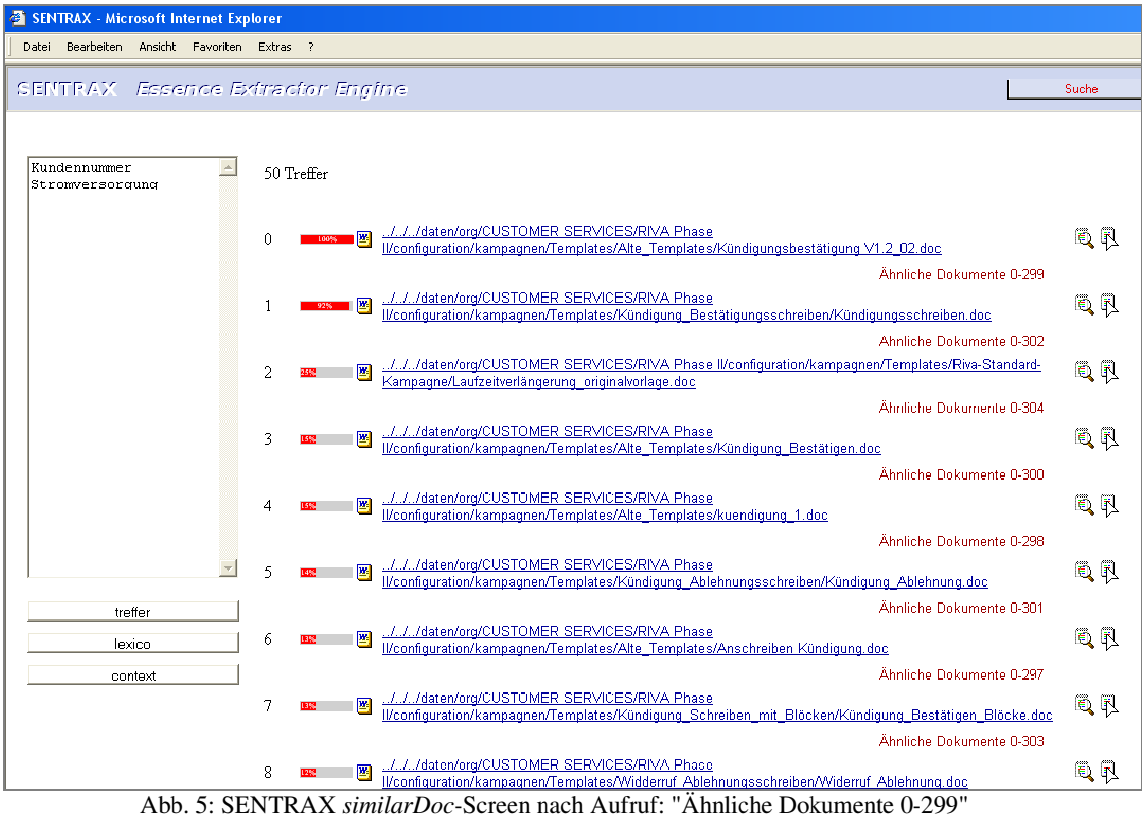

Abb. 5: SENTRAX *similarDoc*-Screen nach Aufruf: "Ähnliche Dokumente 0-299"

Nach unseren Erfahrungen ergeben sich beim intellektuellen Nachprüfen (durch Ansehen der Kandidaten) schon recht große Ähnlichkeiten bei Dokumenten, deren Maßzahl größer als 25% ist. Die Übereinstimmung von 92% im vorliegenden Beispiel rührt von Templates her für ein Kündigungsschreiben (Dokument 0-302) bzw. für eine Kündigungsbestätigung (Dokument 0-299). Bei 100% kann man davon ausgehen, daß es sich praktisch um Doubletten handelt.

Die Suchwörter im Eingabefeld müssen nicht unbedingt in den gefundenen "ähnlichen Dokumenten" vorkommen. Sie bleiben jedoch stehen, damit man einfach (z.B. durch Klick auf *context* oder *treffer*) in die vorherige Recherchephase zurückspringen kann.

(5) Ansichtsoptionen von Dokumenten

Es gibt drei verschiedene Möglichkeiten, sich den Inhalt eines gefundenen Dokuments anzeigen zu lassen.

(i) **Originaldokument.**
Der *treffer*-Bildschirm weist den Pfad mit dem Dateinamen aus. Ein Klick auf diesen Link öffnet das Originaldokument.

(ii) **Dokument im HTML-Format mit Highlight-Funktion**

Durch Klick auf das Symbol  (2. von rechts) öffnet sich eine html-Version des Textinhalts. Der oder die Suchbegriffe sind im Kopf notiert und im Text farblich hervorgehoben (Highlight-Funktion). Durch Scrollen findet man so alle Trefferstellen auf. Alle im Original vorhandenen Links, Grafiken, Bilder etc. sind hier auch vorhanden, können jedoch in der Formatierung und im Layout abweichen.

(iii) **Wie (ii) aber ohne Bilder und Links, dafür mit Sprung zum jeweils nächsten Suchwort.**

Durch Klick auf das Symbol  (ganz rechts) öffnet sich eine Html-Version des Textinhalts. Der oder die Suchbegriffe sind im Kopf notiert und im Text farblich hervorgehoben (Highlight-Funktion). Durch Anklicken des Pfeils in der Kopfzeile bzw. des jeweils "letzten" sichtbaren Trefferworts springt man zum jeweils nächsten. Insofern spart man sich das Scrollen, was besonders bei großen Dokumenten hinderlich ist. Um Konflikte mit bereits bestehenden Links zu verhindern, sind diese aufgehoben.

4 Anwendungsmöglichkeiten und Einsatzfelder

In erster Linie sind die Funktionen der SENTRAX-Engine zum besseren Suchen und Finden in unbekannten, unstrukturierten Textsammlungen konzipiert worden. Die verschiedenen verfügbaren Ähnlichkeitsmaße erlauben aber noch weitergehende Anwendungen. Wenn man sich vorstellt, einen Container mit einer gut definierten Sorte

Texte gefüllt zu haben, z.B. "Kochrezepte" oder "SPAM", dann ließe sich bei einem ankommenden unbekannten Dokument mit den SENTRAX-Funktionen abschätzen, ob es die Merkmale der einen oder der anderen Sorte trägt. So wäre eine automatische Klassifizierung möglich. Im Zusamenhang mit dem Lernen und Arbeiten in netzbasierter Umgebung und mit den Anforderungen des Wissensmanagements ergeben sich wiederum ganz andere Arten der Verwendung dieser Engine. Da die zugrunde liegende Technologie der "Essence Extraction" nicht nur für deutsche Texte sondern auch für viele andere Sprachen wirksam ist, ergeben sich weitere Einsatzmöglichkeiten im Bereich der bi- bzw. multilingualen Recherche. Einige Aspekte dazu sollen kurz angerissen werden.

Wissensmanagement. Sobald ein Unternehmen oder eine Organisation genügend viele Dokumente verfasst und gesammelt hat, gewinnt die Aufgabe, sie zu pflegen, zu sortieren, zu distribuieren, zu archivieren usf. an Bedeutung. Man hat völlig unterschiedliche Sammlungen, wie z.B. Materialien zu Forschungsprojekten, Genehmigungs- und Aufsichtsverfahren, Protokolle, Richtlinien und Gesetzesvorschriften, Handbücher, Präsentationsunterlagen, Archivobjekte. Bei diesen Aufgaben mit textartigen Inhalten kann die SENTRAX-Engine helfen, da unterschiedliche Container definierbar sind, in denen einzeln oder über Auswahlen, aber auch insgesamt gesucht werden kann. Die diversen Funktionen stehen dabei für die jeweiligen Suchanfragen flexibel zur Verfügung, unter Umständen mit spezifischen Parametereinstellungen –z.B. für die Erstellung von Übersichten oder Statistiken. Alle Ausgaben lassen sich auch als Liste darstellen und so wie gewohnt auch automatisch weiterverarbeiten und exportieren.

Kategorisierung. Für die automatische bzw. halbautomatische Kategorisierung von Dokumenten wird in der Literatur gerne die sogenannte SVM-Technik (Support Vector Machine) diskutiert (Kindermann und Leopold [2000]). Die kombinierte Nutzung der SENTRAX Suchfunktionen erlaubt eine alternative Lösungsvariante dieser Problemstellung. Untersuchungen dazu und der Entwurf einer lauffähigen Software wurden erfolgreich auf verschiedene Aufgabenstellungen angewandt. (Vgl. Müller [2002] und Frobese [2006]).

Aus- und Weiterbildung. Definierte Ontologien müssen gut gepflegt werden, insbesondere wenn sie über einen längeren Zeitraum (=mehrere Generationen von Nachwuchsmitarbeitern) gültig, zutreffend, verwendbar sein sollen. Die von der Engine erzeugten *context*-Begriffswolken zu einem Thema oder zu Themenzusammenhängen unterstützen diese Erstellungs- und Pflegearbeiten.

Erzeugt man aus einer geeigneten Dokumentensammlung die zu ausgewählten Fachbegriffen gehörende *context*-Begriffswolke, so kann man diese als „Test“ verwenden, indem man die Probanden zur Bedeutung der Begriffe befragt bzw. deren eventuelle Zusammengehörigkeit erläutern läßt. Ist die zu testende Person bei dem einen oder anderen Fachbegriff unsicher, genügt häufig ein Klick darauf, um mit dem neuen Screen weitere Informationen bzw. Hilfen zu sehen. Gibt man z.B. "statistische Textanalyse" (bei einem passenden Container) ein, dann liefert die SENTRAX die Wortbeziehungen in der Abbildung 6. Es sind mehrere „Fragenkomplexe“ zu sehen, also Begriffsgruppen, die auf Themen referieren, welche in der Ausbildung von Interesse waren. Zum Beispiel: „Was hat statistische Textanalyse mit Rechtschreibfehlerkorrektur zu tun?“ „Was versteht man unter Trigramm

und welche Rolle spielt das (hier)?“ „Welchen Zusammenhang kann Schokolade mit Stimuluswörter haben?“ Usw.

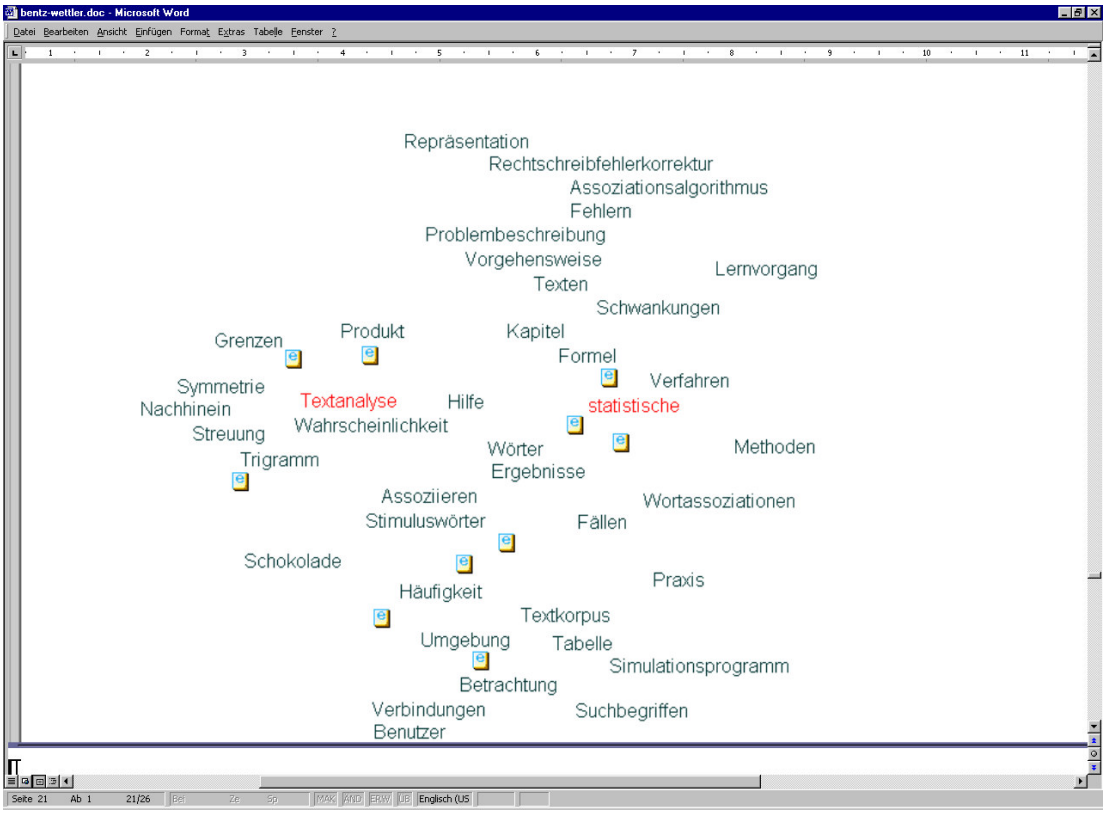

Abb. 6: SENTRAX *context*-Screen zu "statistische Textanalyse", vgl. Bentz [2006]

Wenn man solche Prüfungsscreens nicht aus Forschungsbeiträgen erzeugt, sondern vielmehr eine Sammlung von Lehrbuchtexten zugrunde legt, dann sollten die Gruppierungen noch viel prägnanter werden. Als ein Beispiel dafür wurde eine WBT (Web Based Training)-Einheit zum Thema „Fettstoffwechsel“ indexiert und diese den Teilnehmern der Lernplattform ApoLearn in den Zugriff gestellt. (Vgl. www.apolearn.de bzw. den Link auf der Seite www.imbyte.de). Hier kann der Lernende den Context-Bildschirm zur Orientierung nutzen und von da aus direkt auf das Kapitel zugreifen, in dem das Thema von den Eingabewörtern bzw. den Antwortclustern handelt.

Erfassung von Expertenwissen. Die Ergebnisse der SENTRAX-Contextmap können nicht nur als Antwort auf eine Suchanfrage, sondern als Darstellung eines gewissen Umfeldes zum Thema der eingegebenen Suchbegriffe interpretiert werden. Insofern fungiert die Engine wie ein „passives Gedächtnis“, fördert also sehr effizient die Entscheidung, welche Art von Wissensdomänen oder Wissensaussagen mit ihren charakteristischen Fachbegriffen notiert werden können bzw. müssen. Diese Art Gedächtnis wird aber eine weitaus größere Menge Daten erfassen können als der Mensch es vermag und dabei nie etwas vergessen. Wenn man die Parameter gut einstellt, kann schnell und zuverlässig eine komplette Übersicht aller Vorkommen und Verbindungen von „Konzepten“ in den Dokumenten erhalten werden. Für diese Aufgabe und Anwendung wurde eine Hilfsfunktion *mindshift* implementiert, die die intellektuelle Bearbeitung und Speicherung von Ergebnisbildschirmen erlaubt.

Sprachenübergreifendes Information Retrieval. Im Zusammenhang mit bilingualen Retrievalaufgaben kann man mit Hilfe der SENTRAX-Technologie die Hypothese untersuchen, ob parallele Textkorpora „ähnliche“ Clusterformationen aus den beiden den Sprachen zugeordenten Containern hervorbringen. Man wird dabei zunächst voraussetzen, daß die beiden Sprachen „ähnliche“ Regeln haben, wie etwa Englisch-Deutsch. Die Frage erlaubt eine positive Antwort, wie die Untersuchungen von Suriya Na nhongkai zeigen [2006]. Mit diesem Ansatz kann man sogar inhaltliche Aussagen über Dokumente machen, ohne deren Übersetzung zu haben. Die Ursache liegt darin, daß die Maschine andere Merkmale generiert und verarbeitet als der Mensch, wenn es darum geht, mehrsprachige Dokumente aus einer größeren Sammlung einander zuzuordnen.

Abb. 7: Die SENTRAX *context*-Screens zweier paraller Dokumente (D-E)

5 Die Ähnlichkeitsmaße der SENTRAX Engine

Während die *lexico*-Funktion String-orientiert arbeitet, hauptsächlich auf der Basis von n-Grammen, findet die *context*-Funktion bedeutungsverwandte Begriffe in den Dokumenten. Dazu werden Auftretenshäufigkeiten und nahes Beieinanderstehen von Worten und Wortgruppen in den Texten ausgewertet. Man hat daher oft semantisch verwandte Begriffe in der *Context*Map, wie z.B. *Fusion-Zusammenschluss*, es werden aber auch gänzlich verschiedene Worte dort zusammengebracht, wie z.B. *Ausbildung-Analphabetentum*, weil sie durch die Art ihres Auftretens in den Dokumenten einen Vorgang oder eine Idee repräsentieren. Die Güte dieser Funktion hängt von der Homogenität des Datenmaterials ab. Für normale Texte, die aus ordentlichen Sätzen bestehen, funktioniert die *Context*Map ziemlich gut. Für Wörter, die inhaltlich zusammenhangslos in Tabellen stehen, wie z.B. in Telefonlisten, darf nicht zuviel von dieser Funktion erwartet werden, da der „Kontext“ vom Benutzer nicht zuverlässig interpretiert werden kann.

Die *treffer*-Funktion zeigt alle Dokumente, in denen die Suchwörter enthalten sind mit 100%-Güte an. Im Falle des Fehlens eines oder mehrerer Sucheingaben wird die Ausgabeliste entsprechend modifiziert, so dass ein betroffenes eine Rangabstufung erfährt. Dokumente auf gleicher Stufe werden nach ihrer intern vergebenen ID sortiert. Innerhalb einer festen Prozentgruppe sind also alle Dokumente gleich gut.

Die *similarDoc*-Funktion arbeitet wieder auf den Wörtern (jetzt des gesamten Textes) und sucht entsprechend passende Dokumente zusammen. Auch hier sind alle Treffer auf derselben Prozentstufe gleichermaßen gut. Diese Funktion ist nicht notwendig symmetrisch, was das Empfinden des Benutzers normalerweise nicht stört. Denn auch ohne IR-Systeme kann es vorkommen, daß ein Dokument A bestpassend zum Dokument B ist, B wiederum (weil es vielleicht viel umfangreicher als A ist) besser zu C passt usw.

6 Fazit

Indem sich die SENTRAX von den herkömmlichen Matching Algorithmen und invertierten Listen zur Indexierung löst und eine aufgabenbezogene Mustererkennung mit angepassten Ähnlichkeitsmaßen verwendet, ermöglicht sie eine fehlertolerante und flexible Suche im Datenbestand. Die Visualisierungen über die *Lexico*Map und die *Context*Map bieten eine wertvolle Hilfe beim Erforschen des Korpus und Verfeinern der Suche. Weit über die "normale" Wortsuche hinaus wird in dem neuen Ansatz vor allem das im Information Retrieval vorherrschende Problem der unterschiedlichen Konzeptrepräsentation in Angriff genommen. Verschiedene Autoren haben unterschiedliche Wortwahlen, um bestimmte Vorgänge, Vorfälle, Ideen oder Konzepte zu beschreiben. Die Suchenden wiederum bedienen sich oft noch anderer Begriffe, um Informationen zu diesen Themen aufzufinden. Bei den Suchenden kommt hinzu, dass sie häufig ein „Informationsbedürfnis“ (information need) zu stillen versuchen und vielleicht zunächst gar keine klare Vorstellung davon haben, wie und mit welchen Begriffen sie am besten ans Ziel kommen.

Da nun die SENTRAX Engine die Suchbegriffe innerhalb von Wortwolken visualisiert und ähnliche, d.h. häufig kookkurrierende Begriffe anzeigt, wird dem Nutzer so ein Bild vermittelt, welche Begriffe in seinem Suchfeld eine wichtige Rolle spielen. Er hat damit die Möglichkeit, die Suche schrittweise zu verfeinern und zu präzisieren („Query Reformulation“). Die SENTRAX unterstützt ihn also zielgerichtet beim Auffinden begriffsverwandter Begriffe, die bei invertierten Listen entschlüpfen. (Vgl. Kummer [2006])

Man kann zusammenfassend sagen, daß dem Nutzer mit dieser Technologie ein schnellerer Zugang zu notwendigen Informationen ermöglicht und ein effizienterer Zugriff auf relevante Datenobjekte als bei herkömmlichen Ansätzen gewährt wird.

Literaturverzeichnis

Ackermann, Martin (2000): Statistische Korpusanalyse zum Extrahieren von semantischen Wortrelationen. Dissertation. Hildesheimer Informatik-Berichte 1/2000. Hildesheim: Universität Hildesheim.

Bentz, Hans-Joachim (2002): Lernen und Arbeiten in virtuellen Räumen - Bezüge zu Wissensmanagement, E-HRM & E-Business. In: Handbuch E-Learning: Expertenwissen aus Wissenschaft und Praxis. Hohenstein, Andreas; Wilbers, Karl (Hrsg.). Köln: Deutscher Wirtschaftsdienst.

Bentz, Hans-Joachim (2006): Suchen und Problemlösen in komplexer Umgebung. In: Perspectives on Cognition: A Festschrift for Manfred Wettler. Rapp, Reinhard; Sedlmeier, Peter (Hrsg.). Lengerich: Pabst Science Publishers.

Frobese, Dirk (2006): Suchmethoden der KI angewendet auf elektronische Nachrichten (E-Mails). Universität Hildesheim. Unveröff. Manuskript.

Kindermann, J.; Leopold, E. (2000): Classification of Texts with Support Vector Machines. An Examination of the Efficiency of Kernels and Data-Transformations; 24th Annual Conference of the Gesellschaft für Klassifikation; Passau.

Kummer, Nina (2006): Analyse von Trefferlisten herkömmlicher Suchmaschinen. Universität Hildesheim, Masterarbeit.

Müller, Karen (2002): Automatische Klassifikation von Textdokumenten. Universität Hildesheim. Masterarbeit.

Na nhongkai, Suriya (2006): Untersuchungen zur sprachübergreifenden, bilingualen Suche mit Hilfe der Konzeptnetz-Technologie der SENTRAX-Engine. Universität Hildesheim. Dissertation.

Wettler, M.; Ferber, R.; Rapp, R. (1995). An associative model of word selection in the generation of search queries. Journal of the American Society for Information Science,46 (1995), 685-699.

Klassifikationsaufgaben mit der SENTRAX. Konkreter Fall: Automatische Detektion von SPAM

Dirk T. Frobese

Universität Hildesheim
Institut für Mathematik und Angewandte Informatik
Marienburger Platz 22
31141 Hildesheim
dfrobese@frobese.de

Zusammenfassung

Die Suchfunktionen des SENTRAX-Verfahrens werden für die Klassifizierung von Mails und im Besonderen für die Detektion von SPAM eingesetzt. Die Eigenschaften einer kontextähnlichen Suche und die Fehlertoleranz sollen genutzt werden, um SPAM Nachrichten treffsicher aufzuspüren.

Abstract:

This article introduces the SENTRAX-engine for classification of E-Mails and detection of SPAM.

1 Einleitung

Das Internet und seine Dienste sind mittlerweile fester Bestandteil der Gesellschaft geworden. Ein technisches System, entstanden als Netz zur Unterstützung wissenschaftlicher Arbeiten auf weltweiter Basis, ist auch aus dem privaten und beruflichen Leben nicht mehr wegzudenken. Die Kernfunktionalität ist immer noch die Bereitstellung von Informationen und der Informationsaustausch weltweit, auf Basis von Standards, inhaltlich lediglich einer Selbstkontrolle unterworfen. Als eine der wichtigsten Dienste ist die elektronische Nachricht, die E-Mail oder Mail, anzusehen. Sie hat den traditionellen Briefverkehr dezimiert, und gäbe es nicht andere Angebote aus dem Internet wie zum Beispiel Versandhäuser oder Auktionshäuser (ebay), dann würde die Post kaum noch für den Privatkunden in Erscheinung treten. Laut einer Studie von ARD Online nutzen 78 Prozent der Anwender das Internet hauptsächlich zum Senden und Empfangen von E-Mails. Dies führt aber auch zu anderen, unangenehmen Begleiterscheinungen. Da es so einfach ist E-Mails zu schreiben, bekommt man auch sehr viele davon. Das führt dazu, daß man als aktiver, am Berufsleben aktiv beteiligter Mensch am Tag durchaus an die hundert E-Mails bekommen kann und mehr. Es gibt mittlerweile Ergebnisse aus Untersuchungen, daß der elektronische Mailverkehr nicht mehr ohne Hilfsmittel wie Suchmaschinen und Filter zu bewältigen ist. Hervorzuheben sind die Werbe- oder Müllbotschaften, so genannte SPAM oder Junk-Mails, die häufig auch Computer-Viren mit sich führen. Beispielsweise die Norddeutsche Landes-

bank in Hannover (NORD/LB) erhält bis zu 800 SPAM- Mails pro Stunde [Artikel:NORD/LB 1]. Aber es sind nicht nur SPAMs, sondern auch die Vielzahl der ernsthaften Nach- richten benötigt Zeit zur Bewältigung. Besonders Unternehmen und Dienstleistungsfirmen werden von ihren Kunden verstärktüber den Informationskanal Mail kontaktiert. Eine Veranstaltung wie die Tour de France führt dazu, daß der Berichterstatter und Fernsehsender ARD am Tag ca. 2.000 Mails erhält. Ein weiters Beispiel ist die Bahn [Artikel:xtramind]. Sie erhält pro Jahr ca. 2 Millionen Mail-Anfragen von ihren Kunden. Anbieter von Mail- und Callcenter-Lösungen gehen davon aus, dass eine E-Mail in der Vollkostenrechnung zwischen 3,- bis 5,- Euro je nach Beantwortung und Un- ternehmensstruktur verursachen [Website:www.vera.ag]. Ausgehend von diesen Zahlen ergibt sich ein Kostenpotential von 6 bis 10 Millionen Euro pro Jahr. Daher werden die Verfahren zum Kategorisieren bzw. Klassifizieren der E-Mails immer wichtiger. Nun bietet das von Prof. Bentz an der Universität Hildesheim entwickelte Such- verfahren einen alternativen Ansatz. Es soll nun vorgestellt werden, in wie weit sich das SENTRAX-Verfahren für ein E-Mail Management eignet als die derzeit bekannten und ob sich ein verbesserter Nutzen für die geschilderte Problematik ergibt.

2 Aufgabenstellung

Aus den Ergebnissen des Fachbereich III Informations- und Kommunikationswissenschaften der Universität Hildesheim unter der Leitung von Prof. Bentz ist eine Implementierung assoziativer Suchverfahren (SENTRAX) enstanden. Dieses System wird zur Klassifizierung von Mails eingesetzt. Das SENTRAX-Verfahren kam zur Textklassifikation bisher noch nicht zum Einsatz. Die genaue Funktionsweise wird in Bentz [2006] dargestellt. Die Aufgaben des E-Mail Management fokussieren sich auf folgende Aufgabengebiete:

1. Die Kategorisierung von Nachrichten und damit die kontextabhängige Zuleitung von Mails in vorgegebene Kategorien mit SENTRAX
2. Klassifizierung und Erkennung von Werbenachrichten (SPAM)

Weitergehend soll eine maschinelle Bestimmung möglicher Kategorien betrachtet werden. Die bisherigen Aufgabenstellungen beschränkten sich auf Klassen, die durch repräsentative Mails oder Texte durch den Benutzer in ausreichender Anzahl festgelegt wurden. Nun soll das Verfahren selbst aus einem unstrukturierten Bestand die möglichen Kategorien bestimmen.

3 Vorgehen

Die Durchführung der Untersuchung zur Kategorisierung von exemplarischen Mail-Beständen erfolgt in folgenden Schritten:

1. *Export:* Die E-Mail Bestände werden aus den Client-Anwendungen bereitgestellt.
2. *Normalisierung:* Header und Body der E-Mails werden für den Lernvorgang auf notwendigen Umfang und Darstellung umgeformt, insbesondere die Anhänge.
3. *Lernphase:* Der Großteil des Bestandes wird für den Lernvorgang verwendet.
4. *Detektion:* Auf Basis der vorangegangenen Lernphase werden die übrigen Mails ver-

wendet, um eine Klassifizierung durchzuführen. Inhaltlich sind die E-Mails manuell kategorisiert worden. Natürlich ist es nicht auszuschliessen, daß es fehlerhafte Zuordnungen durch den Anwender gibt. Dies ist aber zu vernachlässigen.

Betrachtet man Aufbau und Inhalt der Junk-Mails, die man so täglich erhält genauer, dann erkennt man immer wiederkehrende Worte bzw. Zusammenhänge zu den Absendern. Es wird daher davon ausgegangen, daß das SENTRAX-Verfahren bei dieser Art von Nachrichten besonders gute Ergebnisse liefert.

Zur Untersuchung werden die folgenden Algorithmen des SENTRAX-Verfahrens eingesetzt:

- Mindmap(Kontext): Darstellung aller exakter Treffer sowie im Kontext verwendete Alternativen
- Lexikomap: Darstellung von Begriffen mit exakter und ähnlicher Schreibweise.
- Trefferliste: Eine Auflistung der Dateien, in der die Suchbegriffe gefunden worden sind, sortiert aufsteigend nach der Häufigkeit der Treffer in Prozent.
- Fehlertolerante Trefferliste: Mit dieser Funktion wird das Verfahren der Lexikomap und der Trefferliste kombiniert. In einer Liste aufsteigend nach Trefferhäufigkeit werden die gefundenen Begriffe mit exakter oder ähnlicher Schreibweise dargestellt.
- Ähnliche: Es wird eine Trefferliste mit Dokumenten ähnlichen Inhalts geliefert.

Alle Verfahren, mit Ausnahme der Lexicomap, liefern eine Liste der Trefferhäufigkeit in den gefundenen Dokumenten. Damit ist die Lexikomap für die Klassifizierung nicht geeignet. An Stelle eines Suchbegriffes wird eine Mail in ihrer vollen Länge als Kombination komplexer Suchbegriffe verwendet. Dabei werden, wie bei der Indizierung auch, nicht relevante Wörter durch eine Stopwortliste entfernt. Anders als bei der Indizierung werden bei der SENTRAX-Suche nicht einzelne Mails indiziert, sondern alle Nachrichten innerhalb einer Kategorie zusammengefasst und dann indiziert. Die oben aufgeführten Verfahren liefern als Ergebnis damit immer eine Liste der Kategorien in der Reihenfolge der größten Trefferhäufigkeit.

4 Ergebnisse

Mit einem exemplarischen Mailbestand wurden die oben genannten Verfahren von SENTRAX angewendet. Durch eine Reihe von Messungen ergibt sich folgende Ergebnistabelle 1.

Die linke Spalte stellt die Anzahl der Kategorien dar, die für eine Messreihe verwendet wurden. Es wurden 12, 9 und 6 Kategorien aus einem Testbestand von Mails verwendet, die sich inhaltlich nur geringüberschnitten. Die folgende Spalte zeigt die Anzahl der gelernten Mails pro Kategorie. Aus den Mails wurden jeweils 50, 100 und 150 Mails als Basis für den Index erlernt. Die Klassifizierung wurde dann mit 10, 20 ,30 und 50 Samples pro Kategorie durchgeführt.

Kat.	Mails		Trefferliste		Fehlertolerant		Kontext		Ähnliche	
	gel.	prob.	alle	Spam	alle	Spam	alle	Spam	alle	Spam
12	50	10	82%	100%	79%	100%	80%	100%	82%	90%
12	50	20	80%	95%	79%	95%	77%	95%	82%	90%
12	50	30	78%	96%	77%	96%	76%	96%	83%	96%
12	50	50	73%	84%	72%	92%	71%	94%	80%	92%
9	50	10	80%	90%	80%	90%	78%	90%	80%	90%
9	50	20	79%	85%	79%	90%	76%	85%	78%	95%
9	50	30	77%	87%	77%	87%	75%	87%	79%	93%
9	50	50	73%	88%	74%	90%	71%	88%	71%	94%
9	100	10	83%	80%	82%	70%	83%	80%	85%	90%
9	100	20	86%	80%	85%	75%	86%	80%	85%	95%
9	100	30	85%	80%	84%	73%	84%	80%	85%	96%
9	100	50	82%	82%	81%	78%	81%	82%	81%	94%
6	50	10	86%	100%	85%	100%	83%	100%	83%	80%
6	50	20	85%	95%	85%	95%	82%	95%	85%	85%
6	50	30	83%	93%	84%	97%	81%	93%	86%	83%
6	50	50	84%	94%	83%	92%	82%	94%	85%	80%
6	100	10	91%	90%	93%	90%	90%	90%	90%	90%
6	100	20	90%	90%	91%	85%	89%	80%	90%	80%
6	100	30	90%	93%	91%	90%	89%	93%	91%	85%
6	100	50	89%	94%	90%	92%	88%	94%	91%	88%
6	150	10	80%	100%	85%	100%	80%	100%	86%	90%
6	150	20	85%	100%	89%	100%	85%	100%	88%	95%
6	150	30	87%	100%	88%	100%	87%	100%	88%	98%
6	150	50	88%	100%	89%	100%	88%	100%	88%	94%

Tabelle 1: Ergebnisse Kategorisierung Mails privat/geschäftlich

Die Ergebnisse der Klassifikation werden in Prozent dargestellt. Das Ergebnis über alle Kategorien wird in der Spalte "alle" dargestellt. Die Quote bzgl. der Detektion von SPAM in der Zeile daneben. Die aufbereiteten Ergebnisse lassen folgende Rückschlüsse zu:

- Optimale Ergebnisse liefern Kategorien, die eine geringe inhaltliche Überschneidung haben
- Eine Verringerung der Kategorien wirkt sich günstig auf das Ergebnis aus
- Schon ab 70 indizierten Mails wird eine Trefferquote von 80% erreicht
- Das SENTRAX-Verfahren liefert Trefferquoten besser als 70% in der Regel 85%
- SPAM-Mails werden besonders gut erkannt. Die Quote ist meist besser als 90%
- die SENTRAX-Funktion ”ähnliche“ liefert die besten Ergebnisse

Damit liefert SENTRAX sehr gute Ergebnisse für die Klassifikation von Mails und ist besonders für die Erkennung von SPAM geeignet.

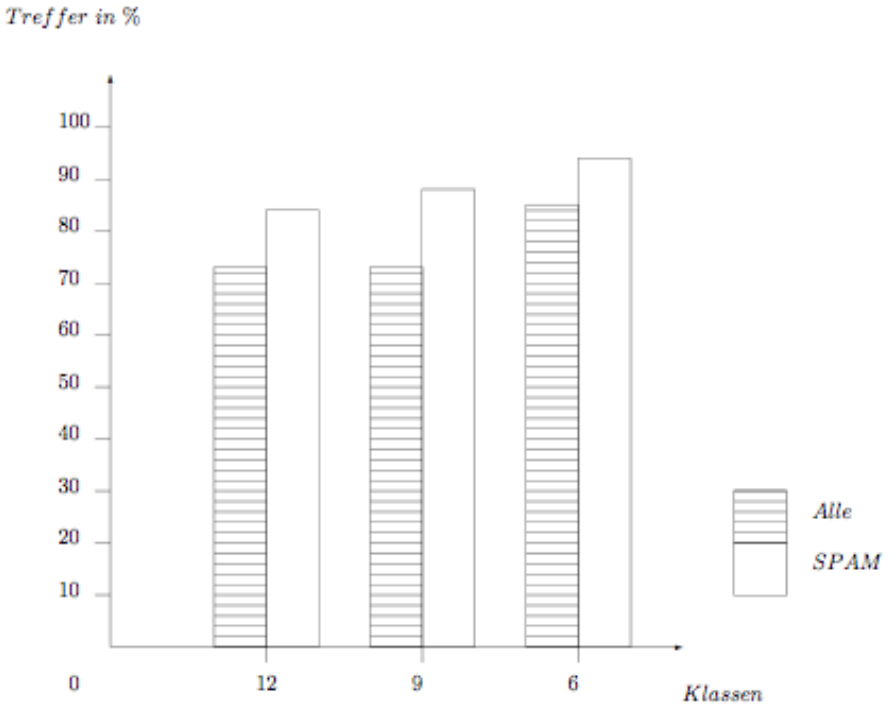

Abbildung 1: Trefferquote SPAM abhängig von der Anzahl der Kategorien

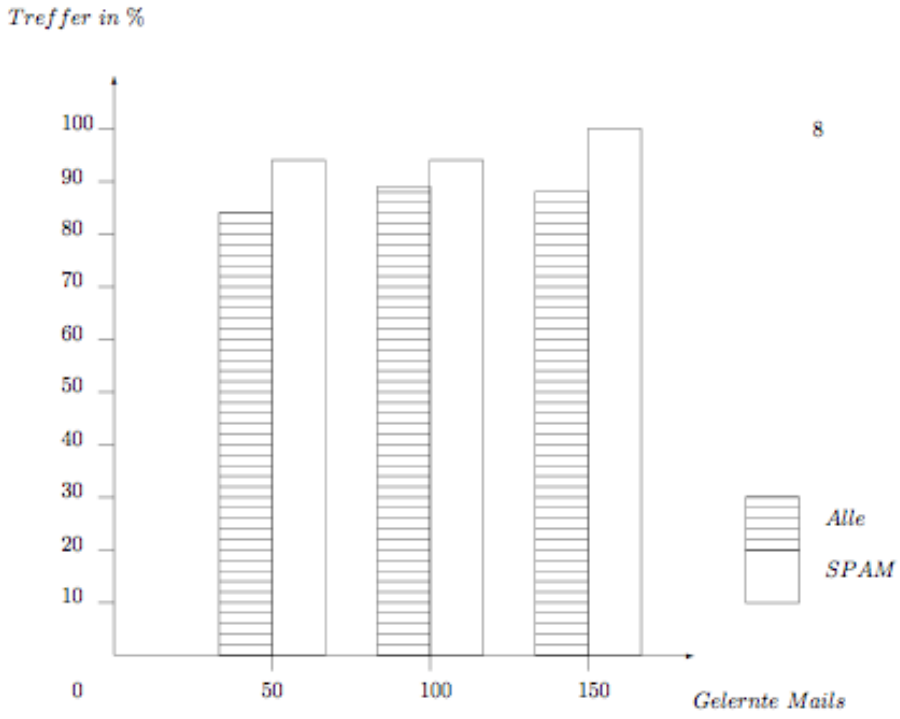

Abbildung 2: Vergleich der Trefferquoten mit SPAM

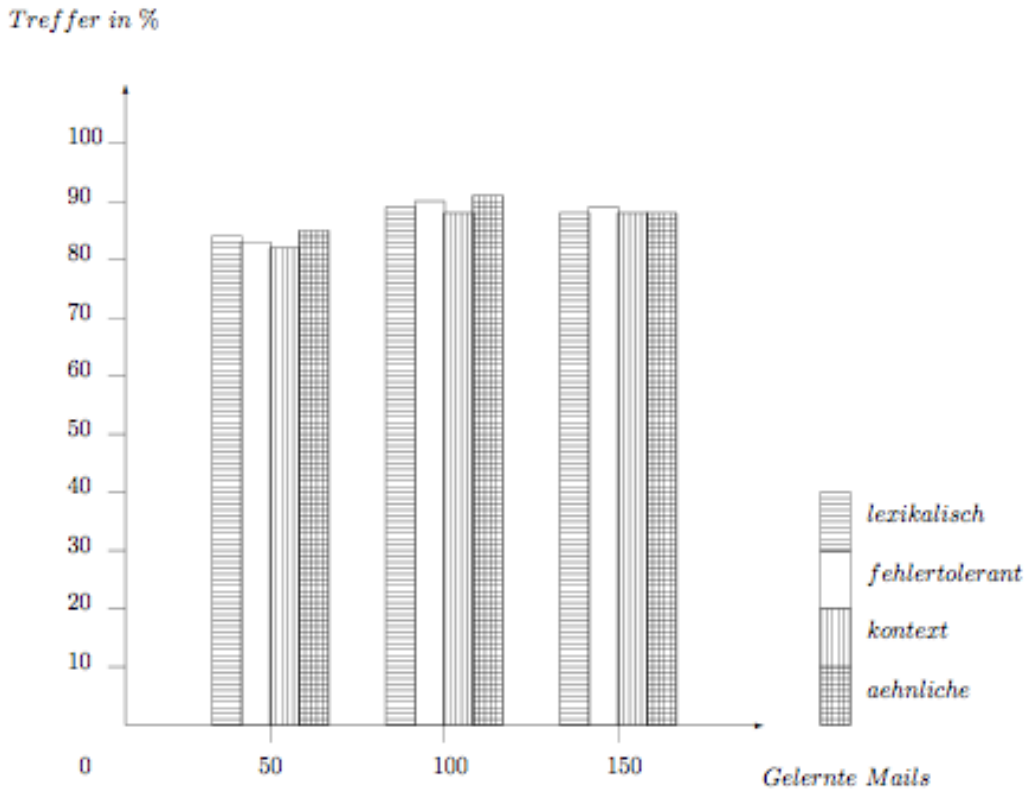

Abbildung 3: Vergleich der Trefferquoten

Literaturverzeichnis

Bentz, Hans-Joachim [2006] Die Suchmaschine SENTRAX. Grundlagen und Anwendungen dieser Neuentwicklung. In diesem Band.

Müller, Karen: Automatische Klassifikation von Textdokumenten, Universität Hildesheim, Dezember 2002

Suriya Na Nhongkai, Hans-Joachim Bentz: Bilinguale Suche mittels Konzeptnetzen, Universität Hildesheim, August 2005

[NORD/LB 1] Artikel Durchblick im Postfach in New Spirit, 1/2004 LITERATUR 11

[Website:www.vera.ag] Hompage der VERA Callcenter Lösung

[Website:www.xtramind.de] Hompage der xtramind E-Mail Management Lösung

[xtramind] xtramind: Success Story DB Dialog Telefonservice GmbH, 8/2004

Bilinguale Suche mit der SENTRAX-Technologie

Myrja Marx, Suriya Na nhongkai

Universität Hildesheim
Institut für Mathematik und Angewandte Informatik
Marienburger Platz 22
31141 Hildesheim
myrja@gmx.de, iamsuriya@yahoo.com

Zusammenfassung

Bei der krosslingualen Suche vermindert eine ungenügende Übereinstimmung mit den Formulierungen im gesuchten Dokument oft die Leistungsfähigkeit der Suche. Hinter der SENTRAX (Essence Extractor Engine) liegen zwei Container (indexierte Dokumente), die für die bilinguale Suche zusammenwirken. Sie entstehen aus der Verarbeitung von nahe zusammenstehenden, bedeutungstragenden Begriffen (Kookkurrenzen) in den zu durchsuchenden Dokumenten und erlauben eine Definition sowie Übertragung von "Konzepten", die zwar durch Worte ausgedrückt oder beschrieben werden, aber eine gewisse Unabhängigkeit von der spezifischen Wortwahl haben. Hierbei kann die Übertragung eines Konzeptes – statt der wortweisen Übersetzung der Anfrage – die Mehrdeutigkeiten entscheidend vermindern, da das Konzept den assoziierten Zusammenhang mit den übersetzten Begriffe bewahrt und die Verbindung zu den Umgebungen in den Texten herstellt. Somit kann sichergestellt werden, dass die dahinter liegenden Dokumente von den gleichen bzw. ähnlichen Themen handeln. Durch grafische Darstellung sind die mit den Suchwörtern assoziierten Begriffe in Ausgangs- und Zielsprache vergleichbar.

Abstract:

A insufficient match of keywords on crosslingual search mostly reduces the capability of information retrieval. For a bilingual search with SENTRAX (Essence Extractor Engine) two containers (“indexes”) are being used, which accrue of word cooccurrences and allow a definition as well as the transmission of concepts. Hereby, the transfer of a concept can minimize the ambiguity of terms because the associated correlation of terms is preserved. Thus, it is ensured that the documents are dealing of similar topics. Due to the graphic display the associated keywords are comparable to source and target language.

1 Einleitung

Die fortschreitende Globalisierung stellt viele Menschen beinahe täglich vor die Herausforderung sich nicht nur mit Informationen in der Muttersprache, sondern in verschiedenen Sprachen auseinanderzusetzen, um sich umfassend zu informieren.

Mit der Zunahme des Datenumfangs erhöht sich jedoch der Bedarf an geeigneter Suchtechnologie. Die meisten Nutzer von Suchmaschinen wissen wenig über Retrievalmethoden und kennen häufig den Gesamtbestand der Dokumentsammlung nicht, in dem sie suchen. Infolgedessen fällt die richtige Anfrage schwer, obgleich sie als wichtiger Schlüssel zur Lösung betrachtet werden muss.

Grootjen und van der Weide haben auf die Schwierigkeit der Anfrageformulierung hingewiesen (Grootjen, F. A.; van der Weide, Th. P., 2002). Zum einen haben sie sich mit der Fragestellung beschäftigt, ob der Benutzer konkret weiß, was er sucht und ob er weiß wie eine optimale Anfrage formuliert werden kann, um die gesuchte Information zu erhalten. Eine gute Anfrage erfordert, dass der Nutzer vorhersieht, welche Ausdrücke in den gesuchten Dokumenten stehen. Das heisst, er muss sich innerhalb von kürzester Zeit einen Überblick über die Dokumentensammlung verschaffen. Unerfahrene Nutzer beispielsweise, haben Probleme herauszufinden, um welche Themen es im entsprechenden Korpus geht. Anhand der Interaktion zwischen dem Nutzer und der Maschine kann der Suchende sein Verständnis für das zu durchsuchende Material langsam aufbauen. Dieser Prozess wird von Spink untersucht (Spink, A., Saracevic, T., 1997). Obwohl die Mensch-Maschine-Interaktion für ungeübte Benutzer am Anfang doppelt schwer ist (bedingt durch die sog. Nutzerbezogene Komplexität), hat sie auch positive Seiten, denn Wahlfreiheiten und Eigengestaltung stellen hohe Anreize und wirken oft motivationsfördernd (vgl. Bentz, H-J., 2005).

2 Bilinguale Suche

Einhergehend mit der Zunahme des Datenumfangs und der Globalisierung kommt die Idee der bilingualen Suche ins Spiel. Die Grundidee des herkömmlichen Ansatzes ist es, zwei oder mehrere Information Retrieval (IR)-Systeme zu verbinden. Statt einen Korpus zu übersetzen, wird die Anfrage ins jeweilige System übertragen und dort separat bearbeitet. Die Ergebnisse der Systeme werden gesammelt und in einer Liste sortiert ausgegeben (siehe Abbildung 1).

Abbildung 1 Context-Liste der SENTRAX (aus: Na nhongkai, S., 2006)

Es wird angenommen, dass das IR-System von genügender Qualität ist. Wenn ein IR-System mit unterschiedlichen sprachlichen Korpora verbunden werden soll, ergeben sich zwei Fragen:

Zum einen stellt sich die Frage, ob die übertragene Anfrage für die fremde Sprache gut geeignet ist. Da eine Übersetzung oft nicht eindeutig ist, kann nicht sicher festgestellt werden, ob die originale Anfrage die gleiche semantische Bedeutung mit der übersetzten Anfrage hat. Falls die Anfragen nicht gleich sind, erhält man möglicherweise nicht das gewünschte Ergebnis.

Hierbei handelt es sich um das Problem der Mehrdeutigkeit von Wörtern. Mehrdeutigkeit betrifft in diesem Kontext Homonymie, Polysemie und Synonymie.
Die Homonymie nimmt Bezug auf Lexeme, die gleich geschrieben sind, aber unterschiedliche Bedeutungen haben („Tau“ = „Seil“ und „Tau“ = „morgendlicher Niederschlag“). Die Polysemie hingegen nimmt Bezug auf die Idee eines einzelnen Lexems mit mehreren relevanten Bedeutungen (z. B. kann "Pferd" für ein Tier, ein Turngerät oder eine Schachfigur stehen).

Das Synonym weist auf die Beziehung zwischen unterschiedlichen Lexemen mit gleicher Bedeutung hin (beispielsweise senkrecht – vertikal, Orange – Apfelsine).
Aufgrund dieser Vielschichtigkeit kann die übersetzte Anfrage den ursprünglichen Sinngehalt der formulierten Anfrage oft nicht bewahren. Eine Möglichkeit die ursprüngliche Bedeutung der formulierten Anfrage nicht zu verlieren, ist bei der SENTRAX dadurch gegeben, dass das gesamte Konzept übertragen wird und nicht nur einzelne Ausdrücke.

Die zweite Frage ist, ob das IR-System auf den beispielsweise englischen Korpus und den Korpus der deutschen Sprache gleich gut passt. Dabei geht es um die Vergleichbarkeit der sprachlichen Eigenschaften. Die englische und die deutsche Sprache zum Beispiel entstammen gleichen Ursprungs, haben sich inzwischen unabhängig voneinander entwickelt. Aus diesem Grund kann nicht eins zu eins übersetzt werden. So etwa ist die deutsche Sprache eine stark flektierende Sprache, die englische hingegen nicht (vgl. Rapp R., 1999). Im Deutschen tauchen oft Komposita auf, z. B. "Finanzamt" oder „Sprachforschung“, im Englischen weniger, z. B. „finance office“ oder „linguistic research“. Aus diesem Grund bedarf es einer gewissen Vorarbeit, um die Symmetrien der Sprachen kenntlich und somit verwertbar zu machen. Eine weitere Möglichkeit besteht darin, ein Wort innerhalb seiner relevanten Umgebung zu betrachten.

Ein neuartiger Ansatz wird wird durch die Suchanwendung SENTRAX aufgezeigt. Hierbei handelt es sich um ein duales IR-Modell, bei dem anhand von Hilfstechniken, wie z. B. Relevanz-Feedback, die Suchanfrage erweitert werden kann, indem Zusatzwörter zu den Suchwörtern anteilig hinzukommen und dem Nutzer eine Interaktion während der Suche anbieten. Besondere Vorteile der SENTRAX sind durch die Lernmöglichkeiten während des Suchprozesses und die Ideen- bzw. Begrifferweiterungen mittels eines Konzeptnetzes gegeben, was bei einer konventionellen Suchmethode nicht vorliegt. Weil der Suchende mit der SENTRAX auf der Konzeptschicht anstatt auf der Wortschicht arbeitet ist es einfacher, weiterführende Zusatzbegriffe auszuwählen, um so das Suchkonzept zu verschärfen und dabei nicht von der Suchrichtung abzulenken (vgl. Na nhongkai, S., 2006).

Das klassische Vorgehen, eine bilinguale Suche aufzubauen, ist die Verknüpfung von zwei IR-Systemen durch eine Übertragungsmethode. Das größte Problem dabei ist die Mehrdeutigkeit der Übersetzung. Ein weiteres Hindernis liegt im Mangel an Ressourcen für die Entwicklung, z. B. der Mangel an elektronisch lesbaren Wörterbüchern für die entsprechende Sprache. Obwohl die Umsetzungssprache solche Mängel umgehen kann, erhöht sich bei wiederholten Übersetzungen trotzdem die Möglichkeit der Mehrdeutigkeit. Bei einer Suchanfrage wird oft ein Konzept (Na nhongkai, S., Bentz, H.-J., 2005) benutzt, um etwas zu definieren, damit die beabsichtigte eindeutige Bedeutung erkannt und verstanden werden kann. So ein Konzept ist in der Regel aus mehreren Eigenschaften zusammengesetzt. Dadurch wird es möglich, dass das gesamte Konzept begriffen wird, obwohl einige Eigenschaften in der Beschreibung fehlen. Anhand dieser Vorgaben kann mittels einer toleranten Konzeptübertragung die Mehrdeutigkeiten bei der bilingualen Suche vermieden werden, weil sich der Kern des Konzeptes nach der Übertragung noch erhält.

Um beim bilingualen Suchschritt die Begriffe der Ausgangsprache in die Zielsprache zu übertragen, stehen zwei Möglichkeiten zur Verfügung. Das elektronisch lesbare deutsch-englische Wörterbuch von der TU-Chemnitz[1] und die Transfermatrix, die der Methode von R. Rapp (1999) entspricht. Dabei wird bei der Übertragung der korpusbasierten Begriffe das gesamte (jeweilige) Suchkonzept bewahrt. Daraufhin werden die Konzeptvergleichsmaße entworfen. Die im Hintergrund stehende Theorie für konzeptionelle Graphabgleichung ist hier an die Vorgaben von M. Montes-y-Gómez, A. López-López und A. F. Gelbukh (2000) angelehnt (vgl. Na nhongkai, S., 2006). Diese Vergleichsmaße dienen zur besseren Auswahl von Suchkonzepten. Die Bausteine des bilingualen Systems werden mit der SENTRAX passend zusammengestellt (siehe Abbildung 2).

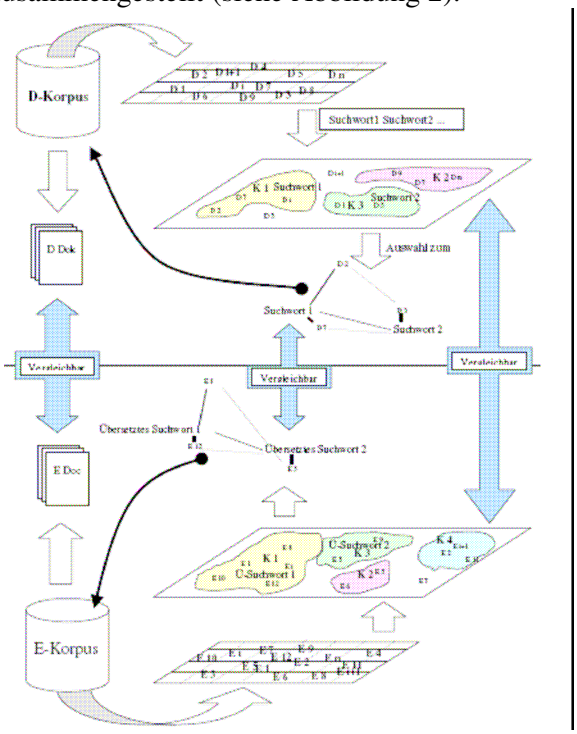

Abbildung 2 Grundidee der bilingualen Suche mit der SENTRAX-Technologie

[1] ftp://ftp.tu-chemnitz.de/pub/Local/urz/ding/de-en/ ,Verifizierungsdatum am 10.10.2006.

3 Mehrsprachige Suche mit der SENTRAX-Technologie

Die SENTRAX ist eine IR-Anwendung mit einer grafischen Mensch-Maschine-Schnittstelle. Sie bietet vier nützliche Funktionen an (vgl. Na Nhongkai, S., Bentz, H.-J., 2005). Anhand der Mensch-Maschine-Schnittstelle kann der Nutzer flexibler arbeiten, als mit einer herkömmlichen IR-Anwendung. Zwei von vier Funktionen werden für Grafikdarstellungen verwendet, wobei Tipp- bzw. Schreibfehler berücksichtigt (LexicoMap, Abbildung 3) und eine andere für die Begriffe zur Erweiterung der Suche verwendet werden (Context-Map).

Abbildung 3 Eingabemaske der SENTRAX mit fehlerhafter Suchanfrage („Konflik“) und korrekter Ausgabe („Nahost-Konflikt“)

Zwei weitere Funktionen, „Treffer“- und „SimilarDoc“ (vgl. Bentz, H.-J., 2006) liefern die Dokumente entsprechend den Suchwörtern als Liste zurück, wobei die jeweilige Prozentangabe einen Zusammenhang zwischen den Dokumenten und den Suchwörtern darstellt. Die Funktion SimilarDoc, kann erst auf der Basis einer erzeugten Trefferliste aktiviert werden. Der Nutzer wählt irgendein Dokument aus und erhält so eine neue Trefferliste, die nun alle ähnlichen Texte im Zusammenhang zum ausgewählten Dokument zeigt. So lassen sich z. B. auch Dubletten oder Fast-Dubletten auffinden.

Die auffälligen Merkmale der SENTRAX sind die grafischen Funktionen. Eine erste Grafikfunktion (LexicoMap) zeigt dem Suchenden diverse Schreibweise als Empfehlung (siehe Abbildung 3). Die im Hintergrund wirksame Technik fußt auf einer SpaCAM. Die SpaCAM-Technik wird in Heitland, M. (1994) beschrieben sowie in M. Hagström, M. (1996). Eine zweite Grafikfunktion ist die zweidimensionale grafische Mensch-Maschine-Schnittstelle, die ContextMap (siehe Abbildung 4, aus: Na nhongkai, S., 2006), die durch die assoziierten Wörter gefüllt wird. Die auf der Grafik-Oberfläche ausgegebenen Wörter haben nicht nur Beziehungen zu der Suchanfrage, sondern auch untereinander. Diese Beziehungen werden in einer kleinen Gruppe klassifiziert, in der eng verwandte Begriffe erfasst werden.

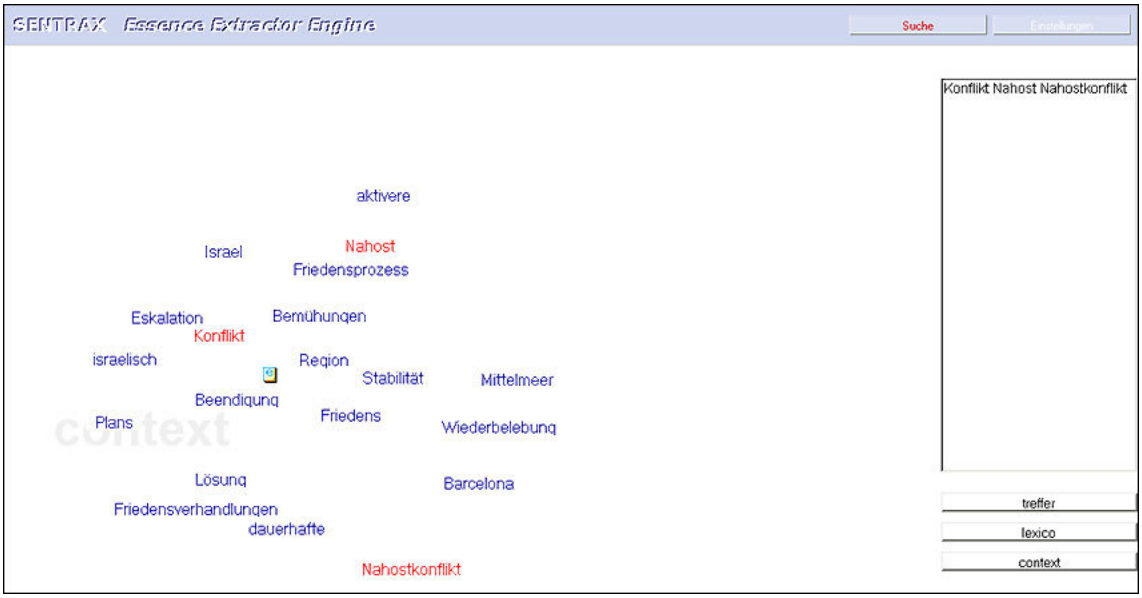

Abbildung 4 ContextMap mit Vorschlägen für eine weiterführende Suche

Der Hintergrund dieser IR-Strategie basiert auf der statistischen Wörterhäufigkeit und der Relation zwischen den Wörtern. Die erste Ordnung bzw. direkte Assoziation wird aus der Häufigkeit berechnet, mit der die Wörter miteinander auftreten. Die Beziehungen der Wörter wie bei Synonymen, Analogien und Antonymen werden durch die zweite Ordnung bzw. indirekte Assoziation ermittelt (vgl. Ackermann, M. 2000). Die Clusterstrategie wird dazu verwendet, um die deutliche Beziehung zwischen den Wörtern auf dem Bildschirm zu zeigen und gleichzeitig in der Gruppe zu klassifizieren. Durch die Mensch-Maschine-Schnittstelle können standardisierte Precision- und Recallwerte nicht gemessen werden.

3.1 Hypothesen

Ausgangspunkt sind zwei Datensammlungen "Deutsch" (D) und "Englisch" (E), die parallel seien. (Eine durch die Parallelität bereits mitgegebene Zuordnung dient lediglich zur späteren Überprüfung unserer Entscheidung, ob das von der Maschine mittels des neuen, schon skizzierten Vorgehens gefundene Dokument das gesuchte Zieldokument in der anderen Sprache ist.). Für D und E werden zunächst unabhängig die SENTRAX-Container erzeugt (vgl. Na Nhongkai, S., Bentz, H.-J., 2005).

Die Vermutung ist, dass bei dieser Datenlage die beiden internen Konzeptnetze eine ähnliche Struktur haben. "Ähnlich" im Sinne der parallelen Dokumentenpaare:

Die Umgebung eines Dokuments in seinem Index "entspricht" der Umgebung seiner Übersetzung im anderen Index. Sollte diese Vorstellung zutreffen, dann müsste die (automatische) Übertragung der einem Dokument hier zugeordneten Wortgruppen ein Cluster von Wörtern dort erzeugen, zu denen das Zieldokument unter allen am besten passt. Als Vorteil bei dieser Methode ist zu erwarten, dass Mehrdeutigkeiten durch die (später vollautomatische) Übersetzung nicht stören, da Bestandteile, die keine Korrespondenzen in den Dokumenten haben, durch den SENTRAX Automatismus unwirksam bleiben. Hierdurch

entsteht eine enorme Reduktion der kombinatorischen Möglichkeiten. Der Umstand, der hier ausgenutzt wird, vergleicht sich mit folgender, "natürlicher" Situation: Wenn man mit einem Menschen spricht, der unsere Sprache nicht besonders gut kann, dann versteht er Gesprächspassagen nicht, in denen Wörter oder Phrasen vorkommen, die ihm fremd sind. Er versteht eben nur das, was in seinem Gehirn eine Korrespondenz in seiner Muttersprache besitzt. Insofern bleibt zuweilen eine große Ausdrucksvielfalt auf unserer Seite nutz- und wirkungslos, da das Verstehenspotenzial auf der anderen Seite sehr eingeschränkt ist.

Die Suchwörter und ihre umgebungsbedingten assoziierten Begriffe, die in der "ContextMap" als einziger Treffer im Container "D" auftreten, werden als Schlüsselwörter bzw. vorgeschlagene Zusatzbegriffe für die Suche im Container "E" verwendet. Eventuelle Mehrdeutigkeiten im Wörterbuch werden unbesorgt übernommen. Die in den Korpora existierende Übersetzung der Schlüsselwörter sollte zum parallelen englischen Dokument entsprechend des vorherigen deutschen Dokuments führen. Diese Vermutung und Methode ist symmetrisch, lässt sich also auch von "E" nach "D" verwenden.

Bei einem großen Container, sollte die Durchmischung der Begriffe auf dem Konzeptnetz gemäß derselben Anfrage möglichst gering werden. Sollte dieses geschehen, könnte das korpusbasierte Semantiknetz mit der ContextMap- Funktion erschaffen werden.

Gibt es kein paralleles Dokument in der Zielsprache entsprechend der Anfrage, sollte das Konzeptnetz zu anderen, ähnlichen Dokumenten führen.

In der Situation, dass der Zielcontainer viel größer oder viel kleiner als der Ausgangscontainer ist, wird die Kookkurrenzhäufigkeit nur gering abweichen, aber die Antwort, nämlich das parallele Dokumentenpaar, sollte noch in dem Dokumententreffer auftauchen. Die bilinguale Suche durch das Konzeptnetz könnte also dennoch funktionieren.

Hat der Zielcontainer mehrere Sprachen als zwei, sollte die bilinguale Suche durch die SENTRAX trotzdem gut funktionieren, weil ein Zusammenhang zwischen Wörtern unterschiedlicher Sprachen nur selten zustande kommen wird.

Für eine nicht-parallele Textsammlung werden die deutschen und englischen Dokumente durch die SENTRAX unabhängig verwaltet. Befänden sich die relevanten Dokumente in beiden Containern, sollten die Konzeptnetze miteinander vergleichbar sein. Wäre der Zusammenhang zwischen den Konzeptnetzen zu schwach, hätten die entsprechenden Dokumente möglicherweise keine Relation zueinander.

3.2 Experimente

Durch die vorhergehend formulierten Hypothesen ergeben sich vier Fälle: (1) der Standardfall, (2) Sonderfälle, (3) die Konzeptnetzänderung und (4) Suche im nicht-parallelen Korpus. Die Verhältnisse im Suchprozess werden für jeden der vier Fälle beobachtet. Die Ergebnisse im Standardfall und in den Sonderfällen bestätigen, dass die bilinguale Suche mittels Konzeptnetzen nicht nur das gesamte Such-Konzept bewahren kann, sondern auch stabil ist. Die Übersetzung wird stets mit dem Online-Wörterbuch http://www.leo.org/ manuell vorgenommen; in der später kommenden Ausbaustufe soll hier auch eine automa-

tische Übersetzung eingeschaltet werden können (vgl. Na Nhongkai, S., Bentz, H.-J., 2005).

Der Suchbedarf beschränkt sich jedoch nicht nur auf eine Richtung wie beim Standardfall Die bilinguale Suche mittels Konzeptnetz funktioniert auch in der Gegenrichtung Englisch → Deutsch (siehe Abbildung 5, aus: Na nhongkai, S., 2006).

Die englische Anfrage in diesem Beispiel ist aus den Wörtern „energy“, „saving“, „ecology“, „environment“ und „research“ zusammengesetzt. Dieses Konzept führt zu dem E-Dokument „ep-01-06-13.txt11“. Die Übersetzung der Anfrage mit Hilfe des Online-Wörterbuchs ist „Energie“, „sparend“, „Ökologie“, „Umwelt“ und „Forschung“. Obwohl das Wort „sparend“ nicht gefunden werden kann, taucht das Wort „einzusparen“ in der deutschen Umgebung auf. Nach der Auswahl des zusätzlichen Attributs „einzusparen“ wird das deutsche parallele Dokument „ep-01-06-13.txt11“ getroffen.

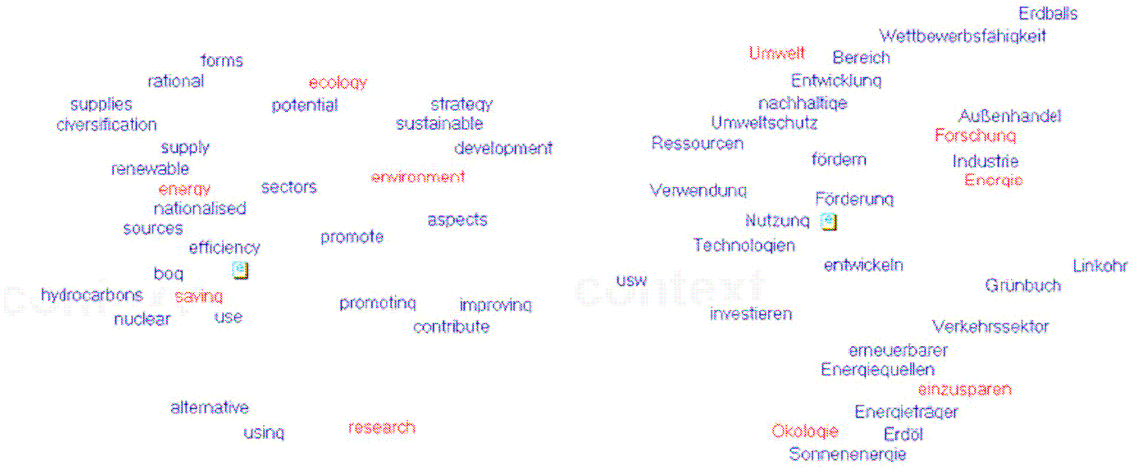

Abbildung 5 Links: ContextMap in der Ausgangsprache Englisch. Rechts: die ContextMap in der Zielsprache Deutsch

Anhand dieses Beispiels wird belegt, dass die Suche in der Gegenrichtung (E→D) ebenfalls funktioniert. Bemerkenswert dabei ist, dass der Zusammenhang der Attribute andere Attribute hervorbringt, wie in diesem Beispiel das Wort „einzusparen“. Die anderen Übersetzungspaare sind natürlich enthalten, z. B. „promote – fördern“, „using (use) – Nutzung (Verwendung)“, „renewable – erneuerbarer“ usw. Das Wort „sectors“ kann vielleicht dem Wort „Bereich“ oder dem Wort „Verkehrssektor“ entsprechen, weil es auf Englisch allein oder mit einem anderen Wort zusammen stehen kann.

Bei den Sonderfällen werden vier Fälle betrachtet: (1) der Zielcontainer ist viel größer als der Ausgangscontainer (2) der Zielcontainer ist kleiner als der Ausgangscontainer (3) das relevante Dokument wird im Zielcontainer entfernt (4) der Zielcontainer wird mit anderen, fremden Texten erweitert. In allen 4 Fällen liefert die SENTRAX zufrieden stellende Ergebnisse, selbst wenn relevante Dokumente aus dem Korpus entfernt werden (vgl. Na Nhongkai, S. 2006, 149 ff.).

Im Falle einer Konzeptnetzänderung wird untersucht, wie das Konzeptnetz bei Größenänderung des Containers verändert wird. Dabei werden zwei Fragestellungen berücksichtigt: Zum einen ob die Entwicklung des Konzeptnetzes bezüglich des deutschen Containers ähnlich zu der Entwicklung des Konzeptnetzes bezüglich des englischen Containers ist und zum anderen, ob sich durch die Vergrößerung des Containers eine Stabilisierung des Konzeptnetzes ergibt.

Die Untersuchungen haben ergeben, dass die Anzahl alter Begriffe bei der Vergrößerung des Containers deutlich höher ist, wenn die neue Textsammlung mit dem vorhergehenden Container zusammen gebildet wurde. Das Änderungsverhalten des deutschen und englischen Konzeptnetzes ist insofern ähnlich, da der inhaltliche Prozess unabhängig von der Sprache abläuft. Dabei ist es irrelevant, ob auf dem englischen oder deutschen Korpus gesucht wird. Wenn eine neue Textsammlung in den Container eingefügt wird, verändert sich die Wortliste hauptsächlich auf den hinteren Rangplätzen. Die neuen Begriffe tauchen in dem Konzeptnetz abhängig davon auf, wie viele Begriffe auf dem Konzeptnetz vom Nutzer eingestellt wurden. Weil das Konzeptnetz von den Termen in der Anfrage abhängig ist, kann man nicht feststellen, wann die Netze in den endgültigen Zustand übergehen. Aber es scheint, dass sich das Konzeptnetz bei stetiger Vergrößerung des Containers langsam entwickelt und stabilisiert (vgl. Na Nhongkai, S. 2006).

Bei der Suche im nicht-parallelen Korpus lässt sich feststellen, dass Konzeptnetze die viele Übersetzungs-vergleichbare Begriffe haben, zu den entsprechenden bilingualen Dokumentenpaaren führen. Durch die sprachliche Vorverarbeitung kann das Problem der unterschiedlichen sprachlichen Nutzungsweise teilweise verhindert werden, indem die Stammform und das Vernachlässigen von unbenötigten Wortarten die Ablenkung durch ungeeignete Begriffe verhindern kann. Als Konsequenz ergibt sich daraus auch eine Änderung der Assoziationsstärke. Dies könnte die sprachliche Symmetrie bringen, die für den Vergleich der Konzeptnetze nötig ist.

5 Fazit

Die Nutzung der zusätzlichen Begriffe aus den relevanten Dokumenten der Ausgangssprache als zusätzliche Übertragungsbegriffe ist eine sinnvolle Methode. Dies kann manuell sowie automatisch erfolgen. Obwohl die manuelle Auswahl einfach und direkt ist, muss der Nutzer viel Zeit aufwenden, um die gefundenen Dokumente zu sichten. Bei der automatischen Auswahl können die ersten *n* gewonnenen Begriffe der Worthäufigkeit oder der Assoziationsstärke nach gemäß der Suchanfrage übernommen werden. Diese Methode kann als „Pseudo-Relevant-Feedback“ bezeichnet werden. Obwohl die Treffer-Funktion der SENTRAX die tolerante Suche mittels der SpaCAM-Technologie ermöglicht, können die getroffenen Dokumente eventuell von dem gesuchten Thema abweichen. Die SENTRAX ist eine Volltextsuchmaschine mittels Musterabgleichung. Wenn ein Dokument viele voneinaner unabhängige Themenbereiche abdeckt, kann die Suche abgelenkt werden. Bei Dokumenten dieser Art können einige Suchbegriffe in einem Thema und andere in einem anderen Themenkomplex vorkommen. Auch wenn die gesuchten Begriffe in einem Dokument weit von einander getrennt stehen, werden sie aufgrund von der Musterabgleichung bei der Volltextsuche als relevant gewertet. Aufgrund der Vielfältigkeit

der Themen und der schwachen Beziehungen durch weit auseinander stehende Wörter sollte ein solches Dokument als unrelevant angesehen werden.

Literaturverzeichnis

Ackermann, Martin (2000): Statistische Korpusanalyse zum Extrahieren von semantischen Wortrelationen. Dissertation. Hildesheimer Informatik-Berichte 1/2000. Hildesheim: Universität Hildesheim.

Bentz, Hans-Joachim (2006): Suchen und Problemlösen in komplexer Umgebung. In: Perspectives on Cognition: A Festschrift for Manfred Wettler. Rapp, Reinhard; Sedlmeier, Peter (Hrsg.). Lengerich: Pabst Science Publishers.

Bentz, Hans-Joachim (2006): Die Suchmaschine SENTRAX. Grundlagen und Anwendungen dieser Neuentwicklung. Universität Hildesheim. Unveröff. Manuskript.

Grootjen, F. A.; van der Weide. Th. P. (2002): Conceptual Relevance Feedback. In: *Proceedingof the 2002 IEEE International Conference on Systems*, Man and Cybernetics, (NLPKE 2002),Tunis, October 2002.

Na nhongkai, Suriya; Bentz, Hans-Joachim (2005): Bilinguale Suche mittels Konzeptnetzen. In: T. Mandl und C. Womser-Hacker (Hrsg.), Effektive Information Retrieval Verfahren in der Praxis: Ausgewählte und erweiterte Beiträge des Vierten Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2005) Hildesheim, 20. Juli 2005. Konstanz: Universitätsverlag [Reihe Schriften zur Informationswissenschaft 45], 2005, 203 – 218.

Na nhongkai, Suriya (2006): Untersuchungen zur sprachübergreifenden, bilingualen Suche mit Hilfe der Konzeptnetz-Technologie der SENTRAX-Engine. Universität Hildesheim. Dissertation.

R. Rapp: Automatic Identification of Word Translations from Unrelated English and German Corpora. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, College Park, Maryland, 519-526, 1999.

Spink, A. und T. Saracevic (1997): Interaction in Information Retrieval: Selection and Effectiveness of Search Terms – Journal of the American Society of Information Science and Technology. 48(8):741-761, 1997.

Entwicklung eines prototypischen Chatbots für die Universitätsbibliothek Hildesheim

Meike Reichle

Universität Hildesheim
Marienburger Platz 22
31141 Hildesheim
meike@alphascorpii.net

Zusammenfassung

Der Beitrag beschreibt die Entwicklung eines Chatbots für die Universitätsbibliothek Hildesheim. Das System antwortet auf Anfragen in getippter natürlicher Sprache. Die Konzeption sowie die Realisierung mit der Artifical Intelligence Markup Language werden besprochen. Eine Evaluierung weist auf eine grundsätzliche Akzeptanz eines derartigen Systems hin.

Abstract:

This article reports the development of a chatbot for the university library in Hildesheim. The System responds to queries in typed natural lanaguage The concept and the implementation with the Artifical Intelligence Markup Language are described. An evaluation hints that such a system may be acceptable for users.

1 Einleitung

Der prototypische Chatbot für die Universitätsbibliothek Hildesheim wurde im Rahmen des Projektseminars *Entwicklung von Chatbots / Avataren*, betreut durch die Dozenten Folker Caroli und Thomas Mandl entwickelt. An der Entwicklung beteiligt waren Studierende der Studiengänge *Internationales Informationsmanagement* und *Informationsmanagement und Informationstechnologie*. Das Projekseminar fand im Sommersemester 2006 statt und umfasste die Konzeption, Planung, Implementation und Evaluierung des Chatbots (Reichle et al. 2006).

2 Konzeption

Nachdem in den ersten Veranstaltungen zunächst ein Überlick über bereits existierende Chatbots, Chatbottypen und -technologien erarbeitet wurde, wurden anschließend verschiedene Anwendungsszenarien im Kurs diskutiert. Kriterien waren hierbei der direkte praktische Bezug innerhalb der Universität, die Realisierbarkeit innerhalb eine Semesters und die

Modellierbarkeit des Anwendungsbereichs. Auf Basis dieser Kriterien wurde ein Informationssystem für die Bibliothek der Universität gewählt, das über häufig auftretende Fragen und die Positionen der einzelnen Fachbereiche innerhalb der Bibliothek Auskunft geben soll. Ein solches Informationssystem soll die Nutzbarkeit der Bibliothek verbessern indem es rund um die Uhr für Fragen zur Verfügung steht und so das Angebot durch die Mitarbeiter der Bibliothek ergänzt. Zusätzlich kann ein Chatbot als nichtmenschlicher Gesprächspartner auch gerade bei häufig gestellten oder scheinbar einfachen Fragen von Nutzen sein, die ein Benutzer den regulären Mitarbeitern vieleicht nicht stellen möchte. Um den Chatbot auch äußerlich in den Kontext der Bibliothek einzubinden soll der fertige Chatbot in die Internetseite der Bibliothek integriert werden.

In der Frage der verwendeten Technologien entschied sich der Kurs für die *Artificial Intelligence Modelling Language* (AIML), einen von Richard Wallace entwickelten XML Dialekt, da in dieser Sprache bereits andere erfolgreiche Chatbot Projekte wie zum Beispiel A.L.I.C.E.[1] implementiert wurden und die Verwendung von AIML auch in Verbindung mit anderen Technologien bereits gut dokumentiert ist (Möbus 2005). Zusätzlich wurden verschiedene Technologien zur Realisierung eines zugehörigen Avatars evaluiert um ein höheres Indentifikationspotential zu schaffen. Hierbei wurde sowohl mit dem javabasierten Avatar-Editor *Edgar*[2] als auch animierten GIFs gearbeitet.

3 Realisierung

Für die weiteren Schritte teilte sich der Kurs in drei Untergruppen auf, die sich einzeln den Bereichen *Inhalte und Dialoge*, *AIML* und *Avatar/Chatbot Generierung* widmeten. Für die wietere Entwicklung wurde ein Meilenstein-Modell gewählt, bei dem die einzelnen Gruppen ihre Arbeit selbst koordinieren und die einzelnen Sitzungen zur Vorstellung der jeweiligen Arbeitsergebnisse der letzten Woche, der Abstimmung zwischen den Gruppen und der Definition neuer Zielvorgaben (Meilensteine) genutzt wurden.

3.1 Inhalte und Dialoge

Aufgabe der Gruppe *Inhalte und Dialoge* war das Erarbeiten der einzelnen Inhalte und der dazugehörigen Dialoge. Zur Zusammenstellung der Inhalte wurde das Inernetangebot der Bibliothek verwendet. Zusätzlich sammelte die Gruppe bei einem Ortstermin am Informationspult der Bibliothek weitere Informationen zu häufig gestellten Fragen und typischen Formulierungen. Ergänzt wurden die gsammelten Informationen schließlich mit selbst erarbeiteten Texten zur Benutzung des Web OPAC Systems der Bibliothek.

Im nächsten Schritt wurden die gesammelten Informationen für die Verwendung mit AIML aufbereitet: AIML arbeitet mit Pattern Matching, das heißt die Eingaben des Nutzers werden mit vorgegebenen Mustern verglichen, denen wiederum entsprechende Antworten zugeordnet sind. Es wurden für die jeweiligen Fragen Muster erarbeitet und entsprechende Antworten formuliert. Schwierig war hierbei vor allem das Finden von

[1] http://www.alicebot.org/
[2] http://i-can-eib.uni-oldenburg.de/ice/buch/

Mustern, die jede mögliche Formulierung einer bestimmten Frage abdecken, ohne auch andere Fragen abzufangen und somit falsche Antworten zu liefern, und das Formulieren von Antworttexten die einerseits möglichst vollständig sind, andererseits aber auch eine schnelle und möglichst genaue Antwort liefern. Die fertig erstellten Muster und Antworttexte wurden der AIML Gruppe zur Überführung in AIML Code und erste Tests übergeben.

3.2 *Artificial Intelligence Markup Language* (AIML)

Die AIML Gruppe überführte die von der Gruppe *Inhalte und Dialoge* erarbeiteten Muster und Inhalte in AIML Code und ergänzte diesen mit bereits fertig erhältlichen Smalltalk Komponenten um den Chatbot durch die zusätzlichen Kommunikationsmöglichkeiten menschenähnlicher und damit auch zugänglicher erscheinen zu lassen. Da die dem Chatbot zu Grunde liegende Infrastruktur parallel von der dritten Gruppe *Avatar/Chatbot Generierung* entwickelt wurde stand sie für Tests noch nicht zur Verfügung. Aus diesem Grunde wurde für die Tests während der Entwicklung der Service Pandorabots.com[3] verwendet, der es erlaubt, kostenlose Chatbots mit eigenen AIML Dateien zu erstellen und im Netz zugänglich zu machen.

3.3 Avatar/Chatbot Generierung

Die Gruppe *Avatar/Chatbot Generierung* realisierte die dem Chatbot zu Grunde liegende Infrastruktur. Diese betseht aus einem Server, der die eigentliche Chatbot Funktionalität zur Verfügung stellt und zusätzlich einem Tomcat Server, der den mit dem Avatareditor Edgar erstellten Avatar unterstützt. Desweiteren war es Aufgabe der Gruppe, einen Avatar zu erstellen, der den entwickelten Chatbot optisch ergänzen sollte. Hierzu wurde der Avatareditor *Edgar* verwendet. Nachdem es sich allerdings mangels entsprechender Dokumentation als unmöglich herausgestellt hatte, die einzelnen Antworten innerhalb des AIML Codes mit entsprechenden Mimiken zu verbinden, wechselte die Gruppe kurz vor dem Ende der Entwicklung zu einem Modell mit animierten GIFs.

4 Evaluation und Fazit

Nach der Fertigstellung des ersten Prototypen wurde eine erste Evaluation vorgenommen. Da die Infrastruktur zu diesem Moment noch nicht entsprechend weit implementiert war, wurde hierzu nochmals der Pandorabots Service verwendet, da hier auch die Protokolle der einzelnen Gespräche einzusehen sind. Die Evaluation wurde von Mitgliedern des Teams *Inhalte und Dialoge* und AIML vorgenommen, als Probanden dienten Studenten der Universität Hildesheim. Die einzelnen Probanden bekamen je einen Chatbot zugeteilt mit dem sie sich über den Zeitraum einer Woche beliebig viel unterhalten konnten. Zusätzlich wurden sie gebeten, einen Fragebogen auszufüllen, der Einschätzungen über den „*Charakter*“ des Chatbots, seine scheinbare Freundlichkeit, Intelligenz und Hilfsbereitschaft enthielt.

[3] http://www.pandorabots.com

Für die Auswertung wurden die Fragebogenergebnisse in Verbindung mit den einzelnen Gesprächsprotokollen herangezogen. Im Ergebnis zeigte sich, dass der Chatbot - obwohl als Informationssystem ausgezeichnet - häufig auch auf sozialer Ebene angesprochen wurde (*Wie geht es Dir heute?*). Außerdem wurden häufig Metafragen (*Was weißt Du? Worüber kannst Du Auskunft geben?*) gestellt, ein Gebiet, dass bei der Entwicklung nicht bedacht worden war.

Bei der Auswertung stellte sich heraus, dass die Benutzer die Freundlichkeit des Chatbots meist positiv bewerteten. Er wurde allerdings auch als eher arrogant eingestuft. Als Grund hierfür wurde das zusätzlich eingebundene Smalltalk Modul identifiziert. Dieses Modul ist hauptsächlich für freie Konversationen im Internet ausgelegt und enthält daher auch zahlreiche eher saloppe Antworten (*Woher soll ich das wissen?*) die im Kontext eines Informationssystems unangebracht sind. Auch die Intelligenz des Chatbots wurde teilweise schlecht bewertet, wenn Fragen *missverstanden*, also den falschen Antworten zugeordnet wurden. In Fällen bei denen die Frage korrekt zugeordnet wurde waren die Benutzer mit Qualität und Ausmaß der Antworten allerdings zufrieden.

Als weitere Maßnahmen wird daher vorgeschlagen, das Smalltalk Modul stark einzuschränken, einige der vorgegebenen Muster zu optimieren und Möglichkeiten zur Selbstauskunft hinzuzufügen. Jenseits dieser Verbesserungsmöglichkeiten ist das Ergebnis allerdings als positiv einzustufen. Der Bibliotheks-Chatbot wurde von den Studenten positiv aufgenommen und häufig wurde auch Interesse an einer Fortsetzung des Projekts geäußert.

Eine ausführliche Beschreibung des Projekts kann in dem nach Ablauf des Projektseminars zusammengestellten Projektbericht gefunden werden. Hier werden die theoretischen Hintergründe eines Chatbots, seine generellen Vor- und Nachteile sowie geeignete Einsatzorte erläutert. Außerdem werden sämtliche in Betracht gezogenen und verwendeten Technologien dargestellt und ihre Tauglichkeit für das Projekt sowie in der Entwicklung aufgetretene Probleme im Detail erläutert. Auch die Ergebnisse der vorgenommenen Evaluation können im Projektbericht ausführlicher nachgelesen werden (Reichle et al. 2006).

Literaturverzeichnis

Möbus, Claus (2005) (Hrsg.): Web-Kommunikation mit OpenSource. Chatbots, Virtuelle Messen, Rich-Media Content. Berlin: Springer.

Reichle, Meike; Matthias Grützner, Stefan Bensch, Zivan Yoash, Marco Blum ,Saskia-Janina Kepp, Bastiaan Scherpenzeel, Jan Maslowski, Dennis Schilling, Marc Ahrens, Kyrylo Streltsov (2006): Avatare und Chatbots. Projektbericht. Informationswissenschaft, Universität Hildesheim.

Entwicklung empirischer Messmethoden zur Validierung der Handlungskompetenz der Piloten

Ruben Weiser

Universität Hildesheim
Informationswissenschaft
Bischofskamp 48
31137 Hildesheim
ruben_weiser@hotmail.com

Zusammenfassung

Im Rahmen des vorligenden Artikels wird ein Bewertungsbogen zur Ermittlung der Handlungskompetenz von Piloten in Zusammenarbeit mit Flugtrainingsexperten der Deutschen Lufthansa AG entwickelt. Die Bewertungen erfolgen in Simulatortests, die vom Luftfahrtbundesamt vorgeschrieben sind und dem Erhalt der Flugzeugmusterberechtigung der Piloten dienen. Zunächst erfolgt die Analyse von Konstrukten und Methoden, die als Kriterien einer Messung zugänglich gemacht werden sollen. Das Ziel der Studie ist es, anhand dieser Konstrukte und Methoden in einer statistischen Analyse der erhobenen Daten Aussagen über das Ausmaß Konstruktvalidität des entwickelten Bewertungsbogens zu formulieren. In diesem Zusammenhang ist zu prüfen, ob die Bewertungen der Fähigkeiten in den klassifizierten Situationen im Flugsimulator generalisierbar sind, oder, ob diese von der spezifischen Situation abhängen.

Abstract

This article deals with the development of an empirical test for monitoring basic competences of commercial pilots. The evaluation sheet is filled out by Lufthansa flight training instructors on a voluntary basis during the period of data collection. The skill ratings are entered during type ratings which are mandatory for all pilots of commercial airplanes due to the regulations of the Luftfahrtbundesamt. An essential part of this study is the development of indicators for operational competence of pilots in emergency situations and their transformation into criteria for measurement. The target of this study is the statistical analysis of the evaluated data to develop statements regarding the construct validity of the created evaluation form. In this context it is to be tested whether or not the results for each competence are independent from the specific method.

1 Einleitung

Sicherheit ist ein wichtiges Thema bei dem zunehmenden Verkehrsaufkommen in der zivilen Luftfahrt. So nimmt etwa der nordamerikanische Flugzeughersteller BOEING an, dass bei der derzeitigen Zuwachsrate des Flugverkehrs die Unfallrate bis zum Jahr 2015 auf zehn große Unglücke pro Jahr ansteigen wird (vgl. NOYES & STARR 1999: S.170).

> „The Human Factor still accounts for the majority of accidents in aviation."
> (STELLING 2004: S. 301)

Die Fähigkeiten der Piloten haben einen entscheidenden Einfluss darauf, ob ein Unfall verhindert werden kann. Daher gehört die regelmäßige Überprüfung der Handlungskompetenz der Piloten zu den wesentlichen Maßnahmen hinsichtlich der Gewährleistung eines sicheren Ablaufes des Flugbetriebs. Die entsprechenden Vorschriften basieren in Europa auf den Vereinbarungen der Joint Aviation Association[1], die in der Bundesrepublik Deutschland durch das Luftfahrtbundesamt umgesetzt werden. Für das Training und die Evaluierung der Fähigkeiten der Cockpitcrews werden regelmäßig Kompetenztests unter realistischen Bedingungen in Flugsimulatoren durchgeführt. Obwohl diese Art der Evaluierung in der Luftfahrtindustrie sehr weit verbreitet ist, bleibt deren Konstruktvalidität allerdings weitgehend ungeprüft (vgl. BEAUBIEN, BAKER, & SALVAGGIO 2004: S. 2). TRUMPOWER et al. haben diesbezüglich in einer exemplarischen Studie zur Überprüfung der Konstruktvalidität der Line Oriented Simulations[2] festgestellt, dass diese Simulatortests offenbar nicht akkurat die spezifischen technischen und kommunikativen Fähigkeiten messen (vgl. TRUMPOWER, JOHNSON & GOLDSMITH 1999: S.1220ff). Da die sicherheitsrelevante Fähigkeit, ein Flugzeug auch in Notfallsituationen sicher fliegen zu können, getestet wird, ist zu postulieren, dass die Gültigkeit der Urteile in solchen Tests so weit wie möglich gewährleistet wird. Zu diesem Zwecke wird im Rahmen dieser Arbeit ein Bewertungsbogen für die Fähigkeiten von Verkehrsflugzeugführern entwickelt, und dessen Konstruktvalidität anhand der Multitrait-Multimethod Methode nach CAMPBELL & FISKE (1959) überprüft[3].

2 Anforderungen an die Piloten der Deutschen Lufthansa AG

Die Inhalte dieses Abschnitts basieren auf den Beobachtungen der OPC/FCL Checks[4] der Deutschen Lufthansa AG. Sie wurden von Experten geprüft und für korrekt befunden.

Die Grundkompetenzen der Flugzeugführer der Deutschen Lufthansa AG werden in drei Kategorien eingeordnet, denen die entsprechenden Fähigkeiten zugeordnet sind. Die Deutsche Lufthansa AG erfüllt mit diesen Kompetenzanforderungen an ihre Piloten alle Kriterien, die in den Joint Aviation Requirements[5] der JAA für den Flugbetrieb[6] und für die

[1] JAA
[2] LOS sind Gate-to-Gate Simulationen, die sämtliche Phasen eines Fluges vom Abflugsgate bis zum Ankunftsgate simulieren.
[3] Zur ausführlichen Lektüre vgl. WEISER (2006)
[4] OPC/FCL Checks (Kompetenztests werden betriebsintern als „Checks" bezeichnet)
[5] JAR

Lizenzierung der Cockpitbesatzungen[7] gefordert werden (vgl. LUFTHANSA ohne Jahr S.1ff.).

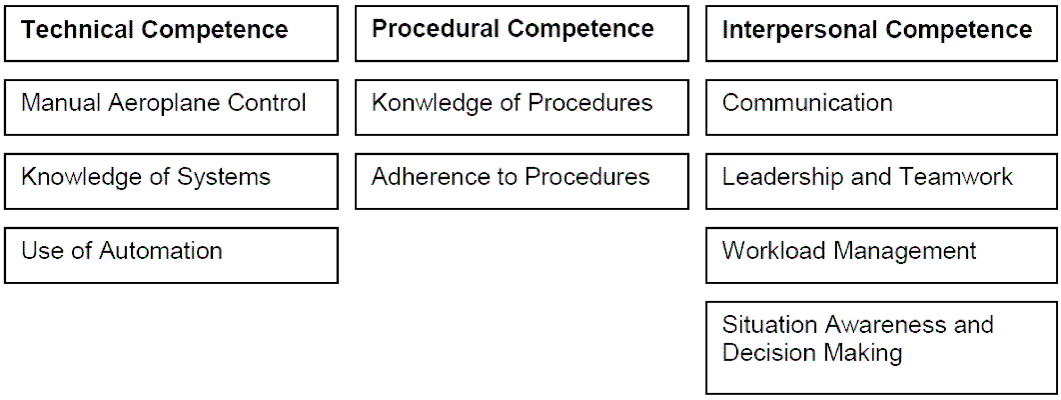

Abb.1 Basiskompetenzen der Piloten

Die Empfehlung der Experten der Deutschen Lufthansa AG lautete, für die Studie Operator Proficiency Checks und Flight Crew Licensing Checks zu verwenden, da in diesen verschiedene Notfallsituationen simuliert werden, was der Notwendigkeit der Klassifizierbarkeit der Bewertungssituationen für die spätere Auswertung am besten entsprach. Die Notfälle sind dementsprechend als Methode zu verstehen, mit der die Fähigkeiten der Piloten geprüft werden.

2.1 Die OPC/FCL Checks

Die Prüfungssituation sieht vor, dass eine Cockpitcrew, grundsätzlich bestehend aus einem Kapitän und einem Copiloten, von einem Check Kapitän in verschiedenen Notsituationen beobachtet wird. Die Aufgabe des Check Kapitäns ist es, die Piloten in den Kategorien interpersonale Kompetenz, prozedurale Kompetenz und technische Kompetenz auf einem standardisierten Formular zu bewerten, um schließlich entscheiden zu können, ob die erforderlichen Kompetenzen vorhanden sind. Für die formulierten Ziele dieser Studie wurde es allerdings erforderlich, eigene Erhebungsbögen zu entwickeln.

2.2 Die Notsituationen

Die Notfälle werden unter anderem in den kritischen Flugphasen Take Off, Steigflug oder Landeanflug simuliert. Einige typische Notsituationen sind ein Triebwerksausfall, ein Feuer am Triebwerk, ein Scherwind, der Ausfall des Radars und der Kollisionskurs mit einem Objekt bei schlechter Sicht. In vielen Fällen gehört es zur vollständigen Behebung des Notfalls, dass das Flugzeug mittels einer Notlandung sicher auf den Boden gebracht wird. Die Notsituationen können teilweise auch in Kombination auftreten und zusätzlich kann Rauch im Cockpit und in der Kabine simuliert werden, was den Zeitdruck zusätzlich erhöht. Zudem können Maßnahmen wie das Eindrücken einer Sicherung zu neuen

[6] JAR-OPS: Joint Aviation Requirements - Operations
[7] JAR-FCL: Joint Aviation Requirements – Flight Crew Licencing

kritischen Momenten wie zum Beispiel zu einem Kabelbrand führen. Es ist allerdings darauf hinzuweisen, dass es in der Wirklichkeit unendlich viele mögliche Verkettungen und Arten von Notfällen gibt. Daher ist es absolut unmöglich, alle Notfälle in Flugsimulatoren zu trainieren. Piloten sind allerdings dazu angehalten, sich in realen Notsituationen an den Simulationen zu orientieren und ihre Handlungen durch Transferdenken an die Umstände anzupassen. Des Weiteren werden viele extreme Situationen wie etwa eine Notwasserung nicht trainiert, da zu viele Faktoren wie etwa der Wellengang unvorhersehbare Variablen sind

3 Empirische Studie: Validierung der Messmethoden

Anhand des vorgestellten Zusammenhangs, in dem die Studie durchgeführt wird, sollen nun inhaltliche Hypothesen formuliert werden, aus denen schließlich statistische Hypothesen zu entwickeln sind, um eine Überprüfung der Konstruktvalidität der entwickelten Messmethode zu ermöglichen. Hypothesenbildung, Auswertung und Interpretation orientieren sich an den Kriterien der Multitrait-Multimethod Methode nach CAMPBELL & FISKE (vgl. CAMPBELL & FISKE 1959: S. 81ff.).

> Konvergente Validität:
> 1. Die konvergenten Validitätskoeffizienten, beziehungsweise deren Mittelwert (Monotrait-Heteromethod Korrelationen), müssen sich signifikant von Null unterscheiden.

Die konvergenten Validitätskoeffizienten sind im Zusammenhang dieser Studie die Korrelationen der Bewertungen derselben Pilotenfähigkeit in zwei verschiedenen Situationen. Drei der vier Kriterien zur Erfüllung der Konstruktvalidität konzentrieren sich auf diese Monotrait-Heteromethod Korrelationen, welche somit eine zentrale Rolle bei der Auswertung spielen. Das zweite und das dritte Kriterium gelten als Indikatoren für die

> Diskriminante Validität:
> 2. Die Monotrait-Heteromethod Korrelationen sollten signifikant größer sein als die Heterotrait-Monomethod Korrelationen.

Die Korrelationen der Bewertungen der gleichen Pilotenfähigkeit in zwei verschiedenen Situationen sollen demnach signifikant größer sein als die Korrelationen der Bewertungen zweier verschiedener Fähigkeiten durch dieselbe Methode. Unterschiede zwischen verschiedenen Fähigkeiten dürfen demzufolge nicht durch die Betrachtung in derselben Situation vermindert werden.

> 3. Die Monotrait-Heteromethod Korrelationen müssen signifikant größer sein als die Heterotrait-Heteromethod Korrelationen. Letztere sollten zudem die geringsten Korrelationen ergeben (vgl. BORTZ & DÖRING 1995: S. 189ff.).

Dieses Kriterium postuliert, dass alle Korrelationen der Bewertungen der gleichen Fähigkeit in zwei verschiedenen Situationen signifikant größer sein sollen als alle Korrelationen der Bewertungen zweier verschiedener Fähigkeiten in zwei verschiedenen Situationen. Die letzteren Heterotrait-Heteromethod Korrelationen haben weder bewertete

Fähigkeit noch Situation gemein. Es ist daher obligatorisch, dass diese Korrelationen in Bezug auf den numerischen Wert die geringsten in der gesamten Matrix sind. Eine Bedingung, die zunächst offensichtlich erscheint, welche aber nicht immer von Tests erfüllt wird (vgl. CAMPBELL & FISKE 1959: S. 82f). Es zeigt sich, dass BORTZ & DÖRING (1995) die Bedingungen der diskriminanten Validität gegenüber CAMPBELL & FISKE (1959) noch verschärfen. In der ersten Niederschrift des Multitrait-Multimethod Ansatzes postulieren CAMPBELL & FISKE, dass jeder Korrelationskoeffizient der Validitätsdiagonalen lediglich signifikant größer sein muss als jene Heterotrait-Korrelationen, welche sich jeweils in derselben Spalte und in derselben Zeile wie der zu überprüfende Validitätskoeffizient befinden. Nach den Kriterien von BORTZ & DÖRING muss jedoch jeder einzelne Validitätdkoeffizient signifikant größer sein als sämtliche Heterotrait-Korrelationen in der Gesamtmatrix (vgl. BORTZ & DÖRING 1995: S. 187ff). Da die gemeinsame Erfüllung der Forderungen für die diskriminante und konvergente Validität Voraussetzungen für die Konstruktvalidität sind, sollte dies unter Betrachtung des vierten Kriteriums untersucht werden. Ein Hinweis darauf sind identische Muster der Korrelationen zwischen verschiedenen Konstrukten in allen Monomethod und Heteromethod Teilmatrizen der Gesamtmatrix.

> 4. Die Rangreihe der Trait-Interkorrelationen sollte in allen Teilmatrizen identisch sein.
>
> „Diese interne ‚Replizierbarkeit' der Rangreihe spricht dafür, dass die ‚wahre' Varianz gemessen wird, bzw. eine echte Korrelationsstruktur zwischen den Traits besteht, die mit den Methoden valide gemessen werden können."
> (BORTZ & DÖRING 1995: S. 190f)

Auch wenn die Messwerte ergeben, dass alle vier Kriterien für die konvergente und diskriminante Validität erfüllt sind, gilt es zu beachten, dass trotzdem Verzerrungen in den Bewertungen auftreten können. Eine letzte Möglichkeit, das Vorliegen einer Konstruktvalidität zu überprüfen, besteht darin die Rangreihen durch Hintergrundwissen plausibel zu machen (vgl. BORTZ & DÖRING 1995: S. 191).

Aus diesen vier Kriterien und den Überlegungen dazu lassen sich nun die statistischen Hypothesen ableiten, welche zur Überprüfung der Konstruktvalidität der zu entwickelnden Messmethoden dienen. Ziel ist es, jeweils die Nullhypothese zugunsten der Alternativhypothese mit einer Irrtumswahrscheinlichkeit $\alpha = 5\%$ zu verwerfen.

> Konvergente Validität:
> H_1: Der Mittelwert der konvergenten Validitätskoeffizienten (Monotrait-Heteromethod Korrelationen) ist signifikant größer als null.
> $\rho A > 0$
> H_0: Der Mittelwert der konvergenten Validitätskoeffizienten (Monotrait-Heteromethod Korrelationen) ist nicht signifikant größer als null.
> $\rho A \leq 0$
>
> Diskrimininante Validität:

H_1: Die Monotrait-Heteromethod Korrelationen ρA sind signifikant größer als die Heterotrait-Heteromethod Korrelationen ρB.
ρA > ρB
H_0: Die Monotrait-Heteromethod Korrelationen ρA sind nicht signifikant größer als die Heterotrait-Heteromethod Korrelationen ρB.
ρA ≤ ρB

Diskrimininante Validität:
H_1: Die Monotrait-Heteromethod Korrelationen ρA sind signifikant größer als die Heterotrait-Monomethod Korrelationen ρC.
ρA >ρC
H_0: Die Monotrait-Heteromethod Korrelationen ρA sind nicht signifikant größer als die Heterotrait-Monomethod Korrelationen ρC.
ρA ≤ρC

Das finale Kriterium lässt sich nicht statistisch, sondern nur durch einen systematischen Vergleich der Rangfolgen überprüfen. Daher kann das entsprechende Hypothesenpaar nur inhaltlich formuliert werden.

Konstruktvalidität:
H_1: Die Rangfolge der Traitkorrelationen ist in den jeweiligen Blöcken der Multitrait-Multimethod Matrix identisch.
H_0: Die Rangfolge der Traitkorrelationen ist in den jeweiligen Blöcken der Multitrait-Multimethod Matrix nicht identisch.

Zur Überprüfung der Hypothesen wurde ein Bewertungsbogen entwickelt, auf dem alle neun Grundkompetenzen von Check Kapitänen in den zwei Notfallsitiationen Engine Failure und Hydraulic Failure zu bewerten waren. Am Ende der Datenerhebung konnten 93 gültige Bewertungsbögen in die Datenauswertung einbezogen werden.

4 Datenauswertung

Die Daten, aus denen die Multitrait-Multimethod Matrix gebildet wird, bestehen zunächst aus den Bewertungen der Fähigkeiten. Die Skalenwerte wurden mit Werten von 4[8] bis 1[9] gemäß den Bewertungen auf den ausgefüllten Bögen in SPSS übertragen. Diese Skalenwerte waren spaltenweise miteinander zu korrelieren, sodass sich eine Matrix mit 18x18 = 324 Korrelationen ergibt. Von diesen sind 18 zu vernachlässigende Autokorrelationen der Variablen mit sich selbst, sodass 306 Korrelationen übrig bleiben. Es ergeben sich 153 relevante Korrelationen der folgenden Gesamtmatrix, die für die Auswertung in Betracht gezogen werden, da jede Korrelation aufgrund der Struktur der Matrix doppelt vertreten ist.

[8] gut
[9] ungenügend

	MAC.EF1	KoS.EF1	UoA.EF1	KoP.EF1	AtP.EF1	C.EF1	LaT.EF1	WM.EF1	SAaDM.EF1	MAC.HF2	KoS.HF2	UoA.HF2	KoP.HF2	AtP.HF2	C.HF2	LaT.HF2	WM.HF2	SAaDM.HF2
MAC.EF1	1,000	0,419	0,124	0,250	0,292	0,077	0,276	0,251	0,421	0,579	0,364	0,173	0,317	0,325	0,224	0,277	0,361	0,379
Sig. 1-seitig		0,000	0,119	0,008	0,002	0,230	0,004	0,008	0,000	0,000	0,000	0,049	0,001	0,001	0,015	0,004	0,000	0,000
KoS.EF1	**0,419	1,000	0,197	0,526	0,524	0,305	0,354	0,223	0,394	0,307	0,601	0,255	0,416	0,434	0,283	0,220	0,168	0,284
Sig. 1-seitig	0,000		0,030	0,000	0,000	0,001	0,000	0,016	0,000	0,001	0,000	0,007	0,000	0,000	0,003	0,017	0,054	0,002
UoA.EF1	0,124	0,196	1,000	0,093	0,249	0,179	0,196	0,200	0,182	0,171	0,291	0,531	0,127	0,274	0,255	0,133	0,112	0,121
Sig. 1-seitig	0,119	0,030		0,188	0,008	0,043	0,030	0,028	0,040	0,050	0,002	0,000	0,113	0,004	0,007	0,102	0,143	0,124
KoP.EF1	**0,250	**0,525	0,093	1,000	0,562	0,329	0,358	0,280	0,404	0,241	0,450	0,190	0,617	0,349	0,329	0,128	0,236	0,469
Sig. 1-seitig	0,008	0,000	0,188		0,000	0,001	0,000	0,006	0,000	0,010	0,000	0,034	0,000	0,000	0,001	0,111	0,011	0,000
AtP.EF1	**0,292	**0,524	**0,249	**0,562	1,000	0,179	0,305	0,362	0,310	0,198	0,409	0,280	0,427	0,750	0,277	0,145	0,316	0,390
Sig. 1-seitig	0,002	0,000	0,008	0,000		0,043	0,002	0,000	0,001	0,028	0,000	0,003	0,000	0,000	0,004	0,083	0,001	0,000
C.EF1	0,077	**0,305	*0,179	**0,329	*0,179	1,000	0,480	0,371	0,224	0,164	0,135	0,102	0,308	0,156	0,609	0,393	0,252	0,172
Sig. 1-seitig	0,230	0,001	0,043	0,001	0,043		0,000	0,000	0,015	0,058	0,098	0,166	0,001	0,068	0,000	0,000	0,007	0,050
LaT.EF1	**0,276	**0,354	*0,196	**0,358	**0,305	**0,480	1,000	0,266	0,363	0,287	0,307	0,199	0,219	0,198	0,351	0,528	0,177	0,220
Sig. 1-seitig	0,004	0,000	0,030	0,000	0,002	0,000		0,005	0,000	0,003	0,001	0,028	0,018	0,029	0,000	0,000	0,044	0,017
WM.EF1	**0,251	*0,223	*0,200	**0,260	**0,362	**0,371	**0,266	1,000	0,263	0,219	0,221	0,193	0,220	0,281	0,259	0,159	0,518	0,298
Sig. 1-seitig	0,008	0,016	0,028	0,006	0,000	0,000	0,005		0,005	0,018	0,017	0,032	0,017	0,003	0,006	0,064	0,000	0,002
SAaDM.EF1	**0,421	**0,394	*0,182	**0,404	**0,310	*0,224	**0,363	**0,263	1,000	0,316	0,394	0,315	0,445	0,208	0,351	0,234	0,262	0,620
Sig. 1-seitig	0,000	0,000	0,040	0,000	0,001	0,015	0,000	0,005		0,001	0,000	0,001	0,000	0,023	0,000	0,012	0,006	0,000
MAC.HF2	**0,579	**0,307	*0,171	*0,241	*0,198	0,164	**0,287	*0,219	**0,316	1,000	0,289	0,325	0,406	0,284	0,289	0,432	0,354	0,358
Sig. 1-seitig	0,000	0,001	0,050	0,010	0,028	0,058	0,003	0,018	0,001		0,003	0,001	0,000	0,003	0,002	0,000	0,000	0,000
KoS.HF2	**0,364	**0,601	**0,291	**0,450	**0,409	0,135	**0,307	*0,221	**0,394	**0,289	1,000	0,367	0,454	0,316	0,361	0,272	0,291	0,372
Sig. 1-seitig	0,000	0,000	0,002	0,000	0,000	0,098	0,001	0,017	0,000	0,003		0,000	0,000	0,001	0,000	0,004	0,002	0,000
UoA.HF2	*0,173	**0,255	**0,531	0,190	**0,280	0,102	*0,199	*0,193	**0,315	**0,325	**0,367	1,000	0,246	0,301	0,333	0,218	0,265	0,383
Sig. 1-seitig	0,049	0,007	0,000	0,034	0,003	0,166	0,028	0,032	0,001	0,001	0,000		0,009	0,002	0,001	0,018	0,005	0,000
KoP.HF2	**0,317	**0,416	0,127	**0,617	**0,427	**0,308	*0,219	*0,220	**0,445	**0,406	**0,454	**0,246	1,000	0,603	0,543	0,363	0,513	0,587
Sig. 1-seitig	0,001	0,000	0,113	0,000	0,000	0,001	0,018	0,017	0,000	0,000	0,000	0,009		0,000	0,000	0,000	0,000	0,000
AtP.HF2	**0,325	**0,434	**0,274	**0,349	**0,750	0,156	*0,198	**0,281	*0,208	**0,284	**0,316	**0,301	**0,603	1,000	0,369	0,306	0,442	0,419
Sig. 1-seitig	0,001	0,000	0,004	0,000	0,000	0,068	0,029	0,003	0,023	0,003	0,001	0,002	0,000		0,000	0,001	0,000	0,000
C.HF2	*0,224	**0,283	**0,255	**0,329	**0,277	**0,609	**0,351	**0,259	**0,351	**0,289	**0,361	**0,333	**0,543	**0,369	1,000	0,521	0,520	0,446
Sig. 1-seitig	0,015	0,003	0,007	0,001	0,004	0,000	0,000	0,006	0,000	0,002	0,000	0,001	0,000	0,000		0,000	0,000	0,000
LaT.HF2	**0,277	*0,220	0,133	0,128	0,145	**0,393	**0,528	0,159	*0,234	**0,432	**0,272	*0,218	**0,363	**0,306	**0,521	1,000	0,353	0,340
Sig. 1-seitig	0,004	0,017	0,102	0,111	0,083	0,000	0,000	0,064	0,012	0,000	0,004	0,018	0,000	0,001	0,000		0,000	0,000
WM.HF2	**0,361	0,188	0,112	*0,236	**0,316	**0,252	*0,177	**0,518	**0,262	**0,354	**0,291	**0,265	**0,513	**0,442	**0,520	**0,353	1,000	0,467
Sig. 1-seitig	0,000	0,054	0,143	0,011	0,001	0,007	0,044	0,000	0,006	0,000	0,002	0,005	0,000	0,000	0,000	0,000		0,000
SAaDM.HF2	**0,379	**0,294	0,121	**0,469	**0,390	*0,172	*0,220	**0,298	**0,620	**0,358	**0,372	**0,383	**0,587	**0,419	**0,446	**0,340	**0,467	1,000
Sig. 1-seitig	0,000	0,002	0,124	0,000	0,000	0,050	0,017	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	

n = 93

** Signifikant auf dem 0,01 Niveau (1seitig)

* Signifikant auf dem 0,05 Niveau (1seitig)

Heterotrait-Monomethod

Heterotrait-Heteromethod

Monotrait-Heteromethod

Monotrait-Monomethod

Abb.2 Multitrait-Multimethod Matrix

5 Datenanalyse

Die Analyse der Korrelationsmatrix umfasst sechs Schritte:

1. Identifikation der verschiedenen Blöcke innerhalb der Matrix
2. Transformation der Korrelationen nach Fisher's z
3. Vergleich der Korrelationskoeffizienten
4. Überprüfung der Signifikanz der Ergebnisse
5. Überprüfung, ob die H_0 verworfen werden kann

Die Überprüfung der Signifikanz erfolgt bei einseitiger Fragestellung und einer Irrtumswahrscheinlichkeit α von 5%.

Nachdem diese Analyse für jede der vier Hypothesen durchgeführt wurde, ergibt sich, dass sich der Mittelwert der konvergenten Validitätskoeffizienten signifikant von null unterscheidet (vgl. BORTZ 1999: S. 775). Ausserdem sind alle Validitätskoeffizienten signifikant größer als die Heterotrait-Heteromethod Korrelationen. Allerdings sind nicht alle Validitätskoeffizienten signifikant größer als die Heterotrait-Monomethod Korrelationen. Auffällig ist aber die Tatsache, dass nur acht der 72 Heterotrait-Monomethod Korrelationen signifikant größer sind als der aus numerischer Sicht geringste, transformierte Validitätskoeffizient. Eine genauere Betrachtung der Rangfolgen der Traitkorrelationen zeigt, dass die Trait-Korrelationen keiner identischen Reihenfolge unterliegen. Wird diesbezüglich allerdings der Umfang von 144 zu vergleichenden Korrelationen in Betracht gezogen, der sich durch die Evaluation von neun Fähigkeiten in zwei Situationen ergibt, so scheint das letzte Kriterium, wie auch die Forderung, dass alle neun Validitätskoeffizienten signifikant größer sein müssen als die 72 Heterotrait-Monomethod Korrelationen aufgrund der hohen Anfälligkeit der Korrelationen für Ausreißer sehr streng zu sein.

6 Diskussion

Die größte Relevanz für die vorliegende Studie hat die Konstruktvalidität, deren Untersuchung anhand der Multitrait-Multimethod Matrix nun nähere Betrachtung erfährt. Die konvergenten Validitätskoeffizienten nehmen generell sehr hohe und signifikante numerische Werte an, wodurch das essentielle Kriterium der konvergenten Validität erfüllt ist. Die beiden verschiedenen Notfallsituationen Engine Failure und Hydraulic Failure erfassen demnach dieselben Konstrukte tendenziell mit einer sehr hohen Übereinstimmung. Eine Fortführung der Untersuchung der Konstruktvalidität war somit gerechtfertigt. Des Wieteren ergibt die Analyse der Multitrait-Multimethod Matrix, dass die Korrelationen in der Validitätsdiagonalen vollständig signifikant größer sind als die Korrelationskoeffizienten des Heterotrait-Heteromethod Blocks, die weder die bewertete Pilotenfähigkeit noch die betrachtete Notfallsituation gemeinsam haben. Mit diesem Ergebnis ist auch eines der zwei Kriterien für diskriminante Validität erfüllt. Ein Hinweis also darauf, dass unterschiedliche Konstrukte in verschiedenen Notfallsituationen differenziert bewertet werden. Diese Ergebnisse lassen in Bezug auf die diskriminante Validität weiterhin vermuten, dass die entwickelte Testmethodik konstruktvalide ist.

Wie sich gezeigt hat, ergibt die statistische Auswertung der Multitrait-Multimethod Matrix, dass alle konvergenten Validitätskoeffizienten signifikant größer als null und vollständig signifikant größer als der numerisch größte Koeffizient der Heterotrait-Heteromethod Korrelation sind. Damit sind die wichtigsten Kriterien zum Nachweis der Konstruktvalidität erfüllt. In Bezug auf die Unterschiede zwischen den Korrelationskoeffizienten ist des Weiteren auffällig, dass alle konvergenten Validitätskoeffizienten signifikant größer sind als ca. 89% der Heterotrait-Monomethod Koeffizienten. Nur knapp 11% der Heterotrait-Monomethod Korrelationen sind signifikant größer als der geringste Wert in der Validitätsdiagonalen. Jeweils nur 5,55% der Heterotrait-Monomethod Korelationen sind größer als die Validitätskoeffizienten auf Rang 7 und Rang 8, nur noch 2,78% sind größer als der Monotrait-Heteromethod Koeffizient auf Rang 6 und nur 1,4% sind größer als der Validitätskoeffizient auf Rang 5. Insgesamt sind alle neun Werte in der Validitätsdiagonalen signifikant größer als 88,88% der Heterotrait-Monomethod Korrelationen. Generell sind die Validitätskoeffizienten signifikant größer als die Heterotrait Korrelationen und es ist folglich eine Tendenz zu erkennen, die eher dafür spricht, dass auch dieses zweite Kriterium der diskriminanten Validität erfüllt werden kann. Zudem gelten geringe Werte in der Validitätsdiagonalen der Multitrait-Multimethod Matrix als Indikator dafür, dass die betrachteten Erhebungsinstrumente nicht valide sind. So sind im Umkehrschluss die durchgehend hohen Werte in der berechneten Validitätsdiagonalen als Hinweis darauf zu interpretieren, dass die Validität der entwickelten Erhebungsinstrumente eher anzunehmen als zu verneinen ist (vgl. CAMPBELL & FISKE 1959: S. 84). CAMPBELL & FISKE äußern zudem bezüglich des Nachweises der diskriminanten Validität:

> „Discriminative validity is not so easily achieved. Just as it is impossible to prove the null hypothesis, or that some object does not exist, so one can never establish that a trait, as measured is differentiated from all other traits. One can only show that this measure of Trait A has little overlap with those measures of B and C, and no dependable generalization beyond B and C can be made.“
> (CAMPBELL & FISKE 1959: S. 103)

Es geht demzufolge weniger darum zu zeigen, dass zwischen zwei unterschiedlichen Fähigkeiten überhaupt kein Zusammenhang besteht, sondern dass dieser, ausgedrückt durch den Korrelationskoeffizienten, in seiner Wertigkeit möglichst gering sein sollte. Die Kriterien zur Erfüllung der Konstruktvalidität eines Tests von CAMPBELL & FISKE und insbesondere die Formulierungen von BORTZ & DÖRING sind aufgrund der nicht zu unterschätzenden Auswirkungen der Ausreißer, also Extremwerte in der Datenverteilung, auf Korrelationen als sehr streng anzusehen (vgl. BORTZ 1999: S.206f.). So kann ein einzelner Ausreißer eine Korrelation der Multitrait-Multimethod Matrix stark verändern und schließlich beim Vergleich der Korrelationsblöcke dafür sorgen, dass ein Kriterium erfüllt wird oder nicht erfüllt wird. Eine Entschärfung der Kriterien wäre daher zu überlegen. Schließlich formulieren auch BORTZ & DÖRING in Bezug auf die Überprüfung der Konstruktvalidität:

> “Der Umstand, dass Testwerte so ausfallen, wie es die aus Theorie und Empirie abgeleiteten Hypothesen vorgeben, kann als Indiz für die Konstruktvalidität des Tests gewertet werden. Eine Konstruktvalidierung ist nur dann erfolgsversprechend, wenn neben dem zu prüfenden Test oder

> Fragebogen ausschließlich gut gesicherte Instrumente verwendet werden und die getesteten Hypothesen Gültigkeit besitzen. Können die Hypothesen nicht bestätigt werden, ist unklar, ob die Validität des Instruments oder die Gültigkeit der Hypothese anzuzweifeln ist. Eine Konstruktvalidierung ist umso überzeugender, je mehr Hypothesen ihre Überprüfung bestehen."
> (BORTZ & DÖRING 1995: S. 186f.)

Die beschriebenen positiven Ergebnisse bezüglich der Konstruktvalidität dürfen nicht verschleiern, dass auch unerwünschte Ergebnisse im Hinblick auf das zweite Kriterium der diskriminanten Validität auftreten. Es lassen sich für fünf der neun Monotrait-Heteromethod Korrelationen signifikant größere Heterotrait-Monomethod Korrelationen isolieren. Insgesamt sind nur vier der neun Validitätskoeffizienten signifikant größer als alle Heterotrait-Korrelationen. Auch das postulierte Kriterium der notwendigen identischen Abfolge der Ränge der Heterotraitkorrelationen kann mit den vorliegenden Daten gemäß der Forderung nicht erfüllt werden.

Die Multitrait-Multimethod Methode ist eine äußerst akribische und detaillierte Variante zur Untersuchung der Konstruktvalidität und in diesem Kontext dürfen die vorliegenden Ergebnisse insgesamt als sehr positiv interpretiert werden. Sie geben daher Grund zur Annahme, dass die Konstruktvalidität des entwickelten Testbogens tendenziell eher angenommen werden kann. Zwar werden die Kriterien nicht vollständig, jedoch in Bezug auf die wesentlichen Aspekte, insbesondere jenes der Konvergenz, erfüllt. Alle konvergenten Validitätskoeffizienten haben einen sehr großen numerischen Wert und für den Großteil der Werte in der Validitätsdiagonalen wird auch das zweite Kriterium der differentiellen Validität erfüllt. Allerdings ist darauf hinzuweisen, dass diese Interpretation der Ergebnisse die Meinung des Autors widerspiegelt.

Die Resultate lassen sich nicht vollständig auf die betriebsinterne Bewertungsmethodik der Deutschen Lufthansa AG in OPC/FCL Checks übertragen, da diese nur eine dichotome Bewertung der Fähigkeiten erlaubt. Da die Wertungen der Prüfer der Deutschen Lufthansa AG jedoch in Bezug auf die betriebsintern verwendeten Kriterien vorgenommen wurden, besteht aufgrund der Ergebnisse der Diskussion insgesamt Grund zu der Annahme, dass auch die Urteile der Ausbilder in den betriebsinternen Verfahren valide sind und somit die tatsächliche Eignung eines Piloten zur souveränen Bewältigung der Notfallsituationen gemessen wird. Basierend darauf wird angenommen, dass die Messmethoden zur Überprüfung der Handlungskompetenz der Piloten bei der Deutschen Lufthansa AG bis auf weiteres als valide betrachtet werden dürfen. Dies lässt sich allerdings im Rahmen dieser Magisterarbeit nicht eindeutig feststellen und bedarf weiterer Untersuchungen.

7 Fazit

Ziel dieser Studie war es, einen Bewertungsbogen für Pilotenfähigkeiten in Notfallsituationen zu entwickeln, diesen im Rahmen einer empirischen Studie bei der Deutschen Lufthansa AG in betriebsinternen Testsituationen durch Ausbilder ausfüllen zu lassen und anschließend das Ausmaß der Konstruktvalidität anhand der Multitrait-Multimethod Matrix zu ermitteln.

Die vorliegende Studie hat Ergebnisse geliefert, die nach Meinung des Autors darauf hinweisen, dass die entwickelten Messinstrumente konstruktvalide sind.

Allerdings ist kein sicherer Schluss auf die Konstruktvalidität des Tests möglich, da nur zwei der vier Nullhypothesen eindeutig verworfen werden konnten. Die Ergebnisse lassen jedoch letztlich die Vermutung, dass die Konstruktvalidität des Tests tendenziell eher hoch zu sein scheint, plausibel erscheinen, da unklar ist, ob die Konstruktvalidität der entwickelten Bewertungsinstrumente anzuzweifeln ist, wenn einzelne Hypothesen nicht bestätigt werden können (vgl. BORTZ & DÖRING 1995: S. 186f.). Grundsätzlich gilt jedoch nach dem Stand der Wissenschaft, dass die Konstruktvalidität mit der Anzahl der Hypothesen steigt, die erfolgreich ihre Überprüfung bestehen. Da zwei der vier Hypothesen ihre Überprüfung vollständig bestehen konnten und keine wissenschaftlich eindeutigen Regeln zur numerischen Überprüfung der Konstruktvalidität existieren, spricht demzufolge sehr viel für die valide Überprüfung der Pilotenfähigkeiten von Verkehrsflugzeugführern der Deutschen Lufthansa AG in Notfallsituationen durch die Messinstrumente.

Das Konzept der Validität und jenes der Reliabilität stehen in dem engen Zusammenhang, dass ein Test bei hoher Reliabilität eine geringe Validität haben kann. Andererseits ist es aber nicht möglich, dass ein Test, dessen Messungen durch Zufallsfaktoren bestimmt werden eine hohe Gültigkeit hat. Es stellt sich dementsprechend die Frage, ob aus der angenommenen Konstruktvalidität des Tests umgekehrt auch das Vorhandensein einer Reliabilität zu folgern ist (vgl. KRECH et al 1992: S.39). Diese Annahme basiert letztlich aber nur auf Vermutungen und soll daher an dieser Stelle als Grundlage für weiterführende Studien formuliert werden, in denen ein Nachweis der Reliabilität erfolgt. Im Rahmen der vorliegenden Studie war es aus aufgrund der geforderten Anonymität nicht möglich, das zweite, wichtige Gütekriterium der Reliabilität eindeutig in einer seiner Facetten zu bestimmen. Bemerkenswert ist des Weiteren der Umstand, dass in der Literatur zwar die Bildung der Multitrait-Multimethod Matrix sowie die Kriterien zur Überprüfung der Konstruktvalidität umfassend beschrieben werden, jedoch zum einen offensichtlich noch immer unklar ist, ab wann ein Messinstrument im Bezug auf die Kriterien von CAMPBELL & FISKE (1959) eindeutig konstruktvalide ist, zum anderen werden die statistischen Formeln zur Berechnung der verschiedenen Signifikanzen nicht geliefert, obwohl dies schon von CAMPBELL & FISKE postuliert wurde.

> „Various statistical treatments for multitrait-multimethod matrices might be developed. We have considered rough tests for the elevation of a value in the validity diagonal above the comparison values in its row and column. Correlations between the Columns for variables measuring the same trait, variance analysis, and factor analysis have been proposed to us. However the statistical development of such statistical methods is beyond the scope of this paper. We believe that such summary statistics are neither necessary nor appropriate at this time."
> (CAMPBELL & FISKE 1959: S.102 f.)

Eine Weiterentwicklung der Kriterien zur Bestimmung der Konstruktvalidität beziehungsweise deren Entschärfung und eine umfassende statistische Anleitung, die auch die notwendigen Signifikanztests beinhaltet, sind wünschenswert und vereinfachten diese ohnehin recht aufwendige Variante der Validierung erheblich.

Solange diesen Vorschlägen nicht entsprochen wird, obliegt es mehr oder weniger der Interpretation und der Auslegung des Wissenschaftlers, ob die überprüften Messinstrumente konstruktvalide sind. Der Anspruch an dieses Forschungsinstrument sollte daher sein, eine eindeutige Vergleichbarkeit der Ergebnisse zu ermöglichen, um Missbrauch zu vermeiden.

Literaturverzeichnis

BAUER, F. (1986): Datenanalyse mit SPSS. Springer Verlag. Berlin et al.

BEAUBIEN, J. M.; BAKER, D. P.; SALVAGGIO, A.M. (2004) Improving the construct validity of line operational simulation ratings: Lessons learned from the Assessment Center. In: THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, 14 (1), S. 1-17

BORTZ, J. (1999): Statistik für Sozialwissenschaftler. Springer Verlag: Berlin, Heidelberg, New York.

BORTZ, J. (1979): Lehrbuch der Statistik für Sozialwissenschaftler. Springer Verlag: Berlin, Heidelberg, New York.

BORTZ. J.; G.A. LIENERT (1998): Kurzgefasste Statistik für die klinische Forschung. Springer Verlag: Berlin et al.

BORTZ. J, DÖRING, N. (1995): Forschungsmethoden und Evaluation. Springer Verlag: Berlin et al.

BROSIUS, F. (1998): SPSS 8.0 Professionelle Statistik unter Windows. MITP-Verlag: Bonn

CAMPBELL, D.T.; FISKE, D.W. (1959): Convergent and Discriminat Validation by the Multitrait-Multimethod Matrix. In: Psychological Bulletin, 56, S. 81-105

DROOG, A. (2004): The Current Status of CRM Training and its Regulations in Europe. In: GOETERS, K.-M. (Hrsg.): Aviation Psychology: Practice and Research. Ashgate Publishing Limited: Aldershot, Burlington.

DROSSE, V.; VOSSEBEIN, U. (Hrsg.) (1999): Statistik Intensivtraining: Repetitorium Wirtschaftswissenschaften. Gabler Verlag: Wiesbaden

DUDEN (2001): Duden Band 5. Fremdwörterbuch. Mannheim et al.: Dudenverlag.

EISSFELDT, H.; GOETERS, K.-M., HOERMANN, MASCHKE, P. & SCHIEWE, A. (1994): DLR Mitteilung 94-09

ERDFELDER, E., MAUSFELD, R., MEISER, T. RUDINGER, G. (Hrsg) (1996): Handbuch Quantitative Methoden. BELTZ PsychologieVerlagsUnion: Weinheim

FLIN, R. (2004): The NOTECHS System. In: GOETERS, K.-M. (Hrsg.): Aviation Psychology: Practice and Research. Ashgate Publishing Limited: Aldershot, Burlington.

FLIN, R.; GOETERS, K.-M.; HORMANN,H.-J.; MARTIN, L.(1998): A Generic Structure of Non-Technical Skills for Training and Assessment. In: Paper presented at the 23rd Conference of the European Association for Aviation Psychology, Vienna, 14th -18th September 1998.

GOETERS, K.-M., MASCHKE, P.; EISSFELDT, H. (2004): Ability Requirements in Core Aviation Professions: Job Analysis of Airline Pilots and Traffic Controllers. In: GOETERS, K.-M. (Hrsg.): Aviation Psychology: Practice and Research. Ashgate Publishing Limited: Aldershot, Burlington.

GOETER, K.-M. (2004): Non-Technical Skills Assessment in Pilot Training: Theory and Practice of the NOTECHS Method. In: GOETERS, K.-M. (Hrsg.): Aviation Psychology: Practice and Research. Ashgate Publishing Limited: Aldershot, Burlington.

HELFRICH, H. (1996): Menschliche Zuverlässigkeit aus sozialpsychologischer Sicht. In: Zeitschrift für Psychologie 204 (1996), S. 15-96

HERRMANN, T; HOFFMANN, M.; KUNAU, G. & LOSER, K.-U, (2004): A modelling method fort he development of groupware applications as socio-technical systems. In Behaviour and Information Technology, March-April 2004, VOL. 23, No. 2, S. 119-135. Taylor and Francis Group.

HOBBST, A.; WILLIAMSON, A. (2002): Skills, rules and knowledge in aircraft maintenance: errors in context. In: ERGONOMICS (2002), VOL 45, NO.4, S. 290-308.

HOEFT, S. & PECENA, Y. (2004): Behaviour-Oriented Evaluation of Aviation Personnel: An Assessment Center Approach. In: GOETERS, K.-M. (Hrsg.): Aviation Psychology: Practice and Research. Ashgate Publishing Limited: Aldershot, Burlington.

HOERMANN, H.-J. (1994): FOR-DEC – A prescriptive method for aeronautical decision making. In: Proceedings of the 21st WEAAP Conference, Dublin

HUNTER, D. R. (2005): Measurement of Hazardous Attitudes among Pilots. In: THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, 15 (1), S. 23-43

JOHANNSON, G. (1993): Mensch-Maschine-Systeme. Springer-Verlag. Berlin et al.

KRECH, D.; CRUTCHFIELD, R.S.; LIVSON, N.; WILSON, A.W.; PARDUCCI, A. (1992): Theoretische Grundlagen und Entwicklungspsychologie In: BENESCH, H. (Hrsg): Grundlagen der Psychologie. Psychologie Verlags Union, Weinheim

KRIZ, J.; LISCH, R. (1988): Methoden Lexikon für Mediziner, Psychologen, Soziologen. Psychologie Verlags Union: München, Weinheim.

LORENZ, B. (2004): Human- Centered Automation: Research and Design Issues. In: GOETERS, K.-M. (Hrsg.): Aviation Psychology: Practice and Research. Ashgate Publishing Limited: Aldershot, Burlington.

LUFTHANSA (ohne Jahr): Basic Competence für Optimum Performance – Competence Criteria for Lufthansa Flight Crew Members. In: internes Arbeitsmaterial

LUFTHANSA (2006): Zukunft Kont erklärt. Lufthansa Artikel.

http://konzern.lufthansa.com/de/html/presse/hintergruende/index.html?c=nachrichten/app/show/de/2003/11/689/HOM&s=0)

verifiziert am 16.08.2006 12:48 Uhr

MOOSBRUGGER, H.; KLUTKY, N. (1987): Regressions- und Varianzanalysen auf der Basis des allgemeinen linearen Modells In: PAWLIK, KURT (Hrsg): Methoden der Psychologie. Verlag Hans Huber. Bern, Stuttgart Toronto

NOYES, J. M.; STARR, A.F. (1999): Civil Aircraft Warning Systems: Future Directions in Information Management and Presentation. In: THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, 10 (2), S. 169-188

O´CONNOR , P. (2004): JAR-TEL Results: Inter-rater Reliabilities, Senistivity and Acceptability of the NOTECHS Method. In: GOETERS, K.-M. (Hrsg.): Aviation Psychology: Practice and Research. Ashgate Publishing Limited: Aldershot, Burlington.

REASON, J. (1992): Menschliches Versagen. Spektrum. Heidelberg, Berlin, Oxford (1994)

ROCHEL, H. (1983): Planung und Auswertung von Untersuchungen im Rahmen des allgemeinen linearen Modells. In: ALBERT, D, PAWLIK, K, STAPF, K.-H., STROEBE, W.(Hrsg.): Lehr- und Forschungstexte Psychologie.

SCHICK, F. (2004): Human/Machine Interfaces for Cooperative Flight Guidance. In: GOETERS, K.-M. (Hrsg.): Aviation Psychology: Practice and Research. Ashgate Publishing Limited: Aldershot, Burlington.

SCHILLING, O. (2001): Grundkurs: Statistik für Psychologen.UTB für Wissenschaft München

SEAMSTER, T. L.; KANKI, B.G. (Hrsg.) (2002): Aviation Information Management. Ashgate Publishing Limited: Aldershot, Burlington.

SHERIDAN, T. B. (1992): Telerobotics, automation, and human supervisory control. The MIT Press Cambridge, Massachusetts, London, England.

STELLING, D. (2004): Psychological Requirements and Examination Guideleines in JAR-FCL3. In: GOETERS, K.-M. (Hrsg.): Aviation Psychology: Practice and Research. Ashgate Publishing Limited: Aldershot, Burlington.

STUMPF, H. (1996): Klassische Testtheorie. IN: ERDFELDER, E.; MAUSFELD, R., MEISER, T. & RUDINGER, G. (Hrsg.) (1996): Handbuch Quantitative Methoden. Psychologie Verlags Union: Weinheim

TRUMPOWER, D. L. et al. (1999). Structural analysis of line-oriented evaluation data. In: R. S. Jensen (Ed.), Proceedings of the 10th International Symposium on Aviation Psychology S. 1220-1223. Columbus: The Ohio State University Press.

WEISER. R. (2006): Entwicklung von empirischen Messmethoden zur Validierung der Handlungskompetenz von Piloten. Magisterarbeit: Stiftung Universität Hildesheim. http://www.uni-hildesheim.de/de/8168.htm

WIEGMANN, D. A.; SHAPPELL, S.A. (1996): A Human Error Approach to Accident Investigation: The Taxonomy of Unsafe Operations. In: THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, 7 (4), S. 269-291

WIEGMANN, D. A., SHAPPELL, S.A. (1995): Human Error Approach Analysis of Postaccident Data: Applying Theoretical Taxonomies of Human Error. In: THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, 7 (1), S. 67-81

WIEGMANN, D.A.; GOH, J. (2001): Visual Flight Rules Into Instrument Meterological Conditions: An Empirical Investigation of the Possible Causes. In: THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, 11 (4), S. 395-379

WIEGMANN, D.A.; SHAPPELL, S.A. (2000): Human Error Perspectives in Aviation. In: THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY, 11 (4), S. 341-357.

WIKIPEDIA (2006): Joint Aviation Authorities erklärt. WIKIPEDIA Artikel. http://de.wikipedia.org/wiki/Joint_Aviation_Authorities. verifiziert am 10.08.2006 21:56 Uhr

WIKIPEDIA (2006): Joint Aviation Requirements erklärt. WIKIPEDIA Artikel. http://de.wikipedia.org/wiki/Joint_Aviation_Authorities verifiziert am 10.08.2006 21:56 Uhr

WIKIPEDIA (2006): Joint Aviation Authorities – Operations erklärt. WIKIPEDIA Artikel. http://de.wikipedia.org/wiki/Joint_Aviation_Authorities verifiziert am 10.08.2006 21:56 Uhr

WIKIPEDIA (2006): Joint Aviation Authorities - Flight Crew Licensing erklärt. WIKIPEDIA Artikel. http://de.wikipedia.org/wiki/Joint_Aviation_Authorities verifiziert am 10.08.2006 21:56 Uhr

WIKIPEDIA (2006): Joint Aviation Requirements - Maintenance erklärt. WIKIPEDIA Artikel. http://de.wikipedia.org/wiki/Joint_Aviation_Authorities verifiziert am 10.08.2006 21:56 Uhr

WIKIPEDIA (2006): Air Transport Pilot License erklärt. WIKIPEDIA Artikel. http://de.wikipedia.org/wiki/Airline_Transport_Pilot_Lincense verifiziert am 16.08.2006 13:29 Uhr

WIKIPEDIA (2006): Pilot erklärt. WIKIPEDIA Artikel. http://de.wikipedia.org/wiki/Pilot verifiziert am 16.08.2006 13:51 Uhr

WOTTAWA, H., THIERAU, M. (1998): Lehrbuch Evaluation. 2. vollst. Überarb. Aufl.: Verlag Hans Huber. Bern, Göttingen, Toronto, Seattle. 1998

Feld-Spezifische Indexierung von Internet-Dokumenten im Rahmen von WebCLEF 2006

Ben Heuwing, Robert Strötgen

Universität Hildesheim
Informationswissenschaft
Marienburger Platz 22
31141 Hildesheim
ben.heuwing@uni-hildesheim.de

Zusammenfassung:

Im Rahmen von WebCLEF 2006 wurde an der Universität Hildesheim mit dem sehr umfangreichen, multilingualen EuroGOV-Korpus experimentiert. Im Vordergrund stand die feldspezifische Indexierung anhand von HTML Strukturelementen. Zusätzlich wurde der Einsatz von Blind Relevance Feedback evaluiert. Wie 2005 wurde ein sprachunabhängiger Indexierungsansatz verwendet. Experimentiert wurde mit dem HTML-Title Element, dem H1 Element und anderen Auszeichnungen, die Text hervorheben. Blind Relevance Feedback wurde für alle Felder außer für das Volltextfeld ‚content' implementiert. Die besten Resultate wurden mit einer starken Gewichtung der HTML-Title und H1 Elemente erreicht und stellten eine geringfügige Verbesserung gegenüber den Ergebnissen aus den letztjährigen Postexperimenten dar. Der Einsatz von Blind Relevance Feedback führte nicht zu Verbesserungen. Für WebCLEF 2006 wurden verbesserte Ergebnisse mit den manuell erstellten Anfragen erreicht, während von den Veranstaltern automatisch erstellte Anfragen zu Ergebnissen führten, die wesentlich unter denen der manuell erstellten lagen. Dies war bei allen teilnehmenden Gruppen der Fall.

Abstract:

For WebCLEF 2006 we experimented with the large, multilingual EuroGOV-Collection. Fieldspecific Indexing using the HTML structure of the web documents was evaluated. In addition, blind relevance feedback was applied in the search process. As in 2005, the experiments were carried out with a language independent indexing strategy. We experimented with HTML title, H1 element and other elements emphasizing text. Blind relevance feedback was implemented for all index fields except for the full content. The best results with the WebCLEF 2005 topics were achieved with a strong weight on the title-element accomplishing a marginal improvement over the best post submission runs for the mixed-monolingual task at WebCLEF 2005. Blind relevance feedback could not yet improve results. For the WebCLEF 2006 topics, improved results were achieved with the manually generated topics, while those automatically generated led to results far below average for all groups participating.

1. Einleitung

Vor dem Hintergrund eines sich erweiternden Europas mit gestiegener Sprachenvielfalt und der Bedeutung von Suchmaschinen als für viele Nutzer wichtigstem Zugangspunkt zum Internet ist die Entwicklung und Evaluierung von Systemen für multilinguales Web-Retrieval von steigender Bedeutung. Der 2005 als Teil der CLEF-Initiative entstandene WebCLEF-Track[1] bietet eine Plattform, um sich mit den spezifischen Herausforderungen zu befassen, die Mehrsprachigkeit und ein sehr großer Korpus an das Information Retrieval stellen. Die Universität Hildesheim hatte nach der erfolgreichen Teilnahme bei WebCLEF 2005 den Anspruch, das entstandene System zu verbessern und zur Evaluierung weiterer IR-Methoden zu nutzen.

Die für das Web-Retrieval typischen Herausforderungen sind die vergleichsweise hohe Heterogenität und der große Umfang der zu durchsuchenden Kollektion. Bei näherer Betrachtung des für WebCLEF erstellten EuroGOV-Korpus werden auch die Herausforderungen der Multilingualität deutlich. Der Korpus umfasst 3,6 Millionen Dokumente in mehr als 25 Sprachen. Zunächst werden daher der EuroGOV-Korpus und die für WebCLEF entwickelten Aufgabenstellungen vorgestellt. Abschnitt 2 gibt dann einen Überblick über die Systeme und Verfahren, die im Kontext von Web Retrieval gute Ergebnisse in verschiedenen Evaluierungsinitiativen gezeigt haben. Auf der Grundlage der an der Universität Hildesheim für den CLEF Ad-hoc Track entwickelten Systeme konnte für WebCLEF 2005 mit einem sprachunabhängigen Indexierungsansatz das beste System für mehrsprachiges Retrieval entwickelt werden (JENSEN ET AL. 2006 und JENSEN 2005a/b). Das System wurde für WebCLEF 2006 in Hinblick auf die Vorverarbeitung der Daten, die Nutzung weiterer HTML-Strukturelemente und die Implementierung einer Blind Relevance Feedback-Funktion zur Anfrageerweiterung weiterentwickelt. In Abschnitt 4 werden die zentralen Elemente des Systems vorgestellt. Danach werden die erzielten Ergebnisse ausgewertet und in Bezug zu den Ergebnissen der anderen Teilnehmer gesetzt. Im letzten Punkt sollen die Ergebnisse bewertet und ein Ausblick auf die Maßnahmen gegeben werden, die für die nächste Teilnahme geplant sind.

2 WebCLEF Organisation

2.1 EuroGOV-Korpus

Für den ersten Durchgang von WebCLEF 2005 wurde aus den Internetseiten der Europäischen Union sowie den Seiten der Regierungseinrichtungen der europäischen Mitgliedsstaaten und Russlands eine Dokumentensammlung erstellt. Eine detaillierte Übersicht geben SIGURBJÖRNSSON ET AL. (2005a). Dieser Korpus wurde 2006 ohne Änderungen wieder verwendet. Der 80GB große EuroGOV-Korpus umfasst ca. 3.6 Mio. Internetseiten. Beim Zusammenstellen des Korpus wurde versucht, ein möglichst vollständiges Abbild des gewählten Ausschnittes des Internets zu erstellen. Dabei stellte es sich aufgrund von unterschiedlichen Namenskonventionen und komplexer Strukturen als

[1] http://ilps.science.uva.nl/WebCLEF/

problematisch heraus, die Seiten der Regierungsinstitutionen korrekt zu identifizieren. Bei der Erstellung wurde daher kein Anspruch auf Vollständigkeit erhoben. Stattdessen wurde versucht jeweils die Seiten der Regierung und die der wichtigsten Ministerien mit aufzunehmen. Gespeichert wurden reine Text/HTML Formate und Rich Document Types wie .doc, .pdf und .ps, jedoch keine Grafiken. Während des Crawling-Vorgangs, bei dem die Seiten eingesammelt wurden, entstanden aus unbekannten Gründen außerdem 70.000 leere Dokumente. Der Korpus enthält mehr als 20 verschiedene Sprachen, wobei die Verteilung der Sprachen eng an die Domains gebunden ist. Besonders interessant ist daher auch die europäische Domain eu.int, da sie Dokumente in allen Sprachen der Mitgliedsstaaten enthält.

Der Korpus besteht aus 157 Dateien in einem XML-ähnlichen Format mit jeweils maximal 25.000 Dokumenten. Zu jedem Dokument gibt es einen Eintrag, der Metadaten und den Inhalt des Dokuments enthält. Vorhandene Metadaten sind neben einer eindeutigen Identifikationsnummer beispielsweise die Internetadresse (URL) des Dokuments und Angaben über den Dokumenttyp, wenn diese durch den ausgebenden Server übertragen wurden. Die eigentlichen Dokumente befinden sich in Form von Text innerhalb eines speziellen XML-Elements, dem CDATA-Element. Innerhalb dieses Elements können beliebige andere XML-Elemente auftauchen, ohne als solche verarbeitet zu werden. Dies würde unweigerlich zu Fehlermeldungen führen, da die Internetsprache HTML in hohem Maße XML-Charakteristiken aufweist und dabei verschiedenste Fehlerquellen enthält. Nachteilig war, dass so in den CDATA-Elementen andere CDATA-Elemente aus den Internetdokumenten auftraten. Diese Verschachtelung von CDATA-Elementen ist in XML jedoch nicht zulässig. Dies ist einer der Gründe, warum der Korpus nur XML-ähnlich ist, das Format des Korpus also nicht wohlgeformtes XML ist. Ein weiterer Grund hierfür ist, dass die URLs, die als Metadaten angeben wurden, Sonderzeichen enthalten, die nicht XML konform sind. Problematisch ist auch die Verarbeitung der Rich Document Types, deren binäre Bestandteile ebenfalls in Text umgewandelt wurden, wodurch im Korpus sinnlose Zeichenfolgen entstanden sind, welche auch Zeichen enthalten, die in einem XML-Dokument nicht auftauchen dürfen. Den Teilnehmern wurden Listen zur Verfügung gestellt, welche die leeren Dokumente und alle Dokumente der Typen .doc und .pdf identifizieren und mit denen diese dann herausgefiltert werden können, da diese Dokumente für die gestellten Aufgaben nicht relevant waren. Eine weitere Herausforderung für den Prozess der Vorverarbeitung sind die zahlreichen unterschiedlichen Zeichenkodierungen der Textdokumente.

Der Korpus ist insofern in Bezug auf die enthaltenen Sprachen, die Dokumenttypen und die Zeichenkodierungen, stark heterogen. Dies entspricht den realen Gegebenheiten im Internet und der Aufgabenstellung einer Internetsuche.

2.2 Topics und Topic-Erstellung

Die als Topics bereitgestellten Suchanfragen für WebCLEF 2006, auf deren Grundlage die teilnehmenden Systeme verglichen wurden, bestanden aus 319 manuell erstellten Anfragen (124 neu erstellte und 195 der für WebCLEF 2005 erstellten Topics) und 1620 automatisch erstellten. Die manuellen Topics umfassen insgesamt 11 Sprachen, die automatischen repräsentieren dagegen eher die Sprachvielfalt des Korpus, da sie mit Hilfe von Dokumenten aus allen Domains erstellt wurden (BALOG ET AL. 2006a). Das für die Erstellung der automatischen Topics verwendete Verfahren wurde im Nachhinein bekannt gegeben.

Dabei wurden aus zufällig ausgewählten Zieldokumenten über ein probabilistisches Modell Anfragen aus einem oder zwei Termen erstellt. Beim automatischen Ablauf der Erstellung sollte der Suchvorgang eines echten Nutzers simuliert werden. Daher wurde zusätzlich auch ein gewisser Anteil an Fehlern (Noise) eingebracht (BALOG ET AL. 2006a), was teilweise zu gemischtsprachigen Topcis oder auf andere Art weniger realistischen Anfragen führte, etwa Topic WC0600346 „*bundesministerium der marginalspalte*“.

Die Topics wurden in einem XML-Format zur Verfügung gestellt (vgl. Abb. 1) und enthielten Metadaten zum Topic, wie die Sprache, eine englische Übersetzung, die Zieldomain und ein Nutzerprofil desjenigen, der das Topic erstellt hatte (in Hinblick auf bevorzugte Sprachen). Die Übersetzung und Angaben zum Nutzerprofil enthielten nur die manuell erstellten Topics. Die Metadaten konnten für die Suche eingesetzt werden, wobei dies jeweils angegeben werden musste.

```
- <topic>
    <num>WC0600001</num>
    <title>Arbeiten Deutschen Bundestag</title>
  - <metadata>
    - <topicprofile>
        <language language="DE" />
        <translation language="EN">Job opportunities German Bundestag</translation>
      </topicprofile>
    - <targetprofile>
        <language language="DE" />
        <domain domain="de" />
      </targetprofile>
    - <userprofile>
        <native language="NL" />
        <active language="EN" />
        <active language="DE" />
        <countryofbirth country="NL" />
        <countryofresidence country="NL" />
      </userprofile>
    </metadata>
  </topic>
```

Abb. 1: WebCLEF 2006-Topic

2.3 WebCLEF-Tasks und Bewertung

Im Mittelpunkt der Evaluierung von WebCLEF stehen zwei Bereiche, die typische Aufgaben einer mehrsprachigen Suchmaschine repräsentieren, der *Mixed Monolingual Task* und der *Multilingual Task*. Für beide werden dieselben Topics eingesetzt. Beim Mixed Monolingual Task sollen Ergebnisse jeweils nur in der Sprache geliefert werden, in der die Suchanfrage formuliert wurde. Im multilingualen Task sollen zu jeder Anfrage relevante Ergebnisse in allen Sprachen gefunden werden. Zu jeder Anfrage wird dabei bereits eine englische Übersetzung angegeben. Als relevant gewertet werden dann neben dem ursprünglichen Zieldokument auch ähnliche Dokumente in den verschiedenen Sprachen. Vorstellbare Anwendungsszenarien für eine solche sprachübergreifende Suche sind etwa der Vergleich von Gesetzgebungen verschiedener Länder, Migranten, die Informationen über ein Land suchen, oder Informationssuche im Vorfeld der Eröffnung einer Firma im Ausland (DE RIJKE & SANTOS 2005). Dabei wird vorausgesetzt, dass der Nutzer über aktive oder passive Sprachkenntnisse in mehreren Zielsprachen verfügt. Da die Ergebnisse des multilingualen Tasks von WebCLEF 2005 insgesamt jedoch wenig erfolgsversprechend

waren, sprachen sich viele der Teilnehmer gegen eine Wiederaufnahme aus. Bei WebCLEF 2006 wurde daher nur der Mixed Monolingual Task durchgeführt. Es stand den Teilnehmern jedoch offen, zur Evaluierungszwecken auch Ergebnisse für den multilingualen Task einzureichen.

Aufgrund eingeschränkter Ressourcen für Topic-Erstellung und Relevanzbewertung wurde für die WebCLEF-Tasks das Prinzip der known-item-Suche (SIGURBJÖRNSSON ET AL. 2005b) gewählt. Dabei wird angenommen, dass ein Nutzer mit seiner Suchanfrage das Ziel verfolgt, eine bestimmte, bereits bekannte Seite wieder zu finden. Dabei sollte entweder die Homepage, also die Einstiegsseite einer Website, oder eine bestimmte Seite innerhalb einer Website gefunden werden (Homepage Topics bzw. Named Page Topics). Zu welcher Gruppe ein Topic gehört war zum Zeitpunkt der Suche nicht bekannt. Diese Aufgabenstellung ist typisch für die Nutzung von Internetsuchmaschinen. Das *known-item* Prinzip vereinfacht aber auch die Bewertung, da zu jedem Topic nur ein gültiges Dokument existiert, welches schon bei der Erstellung des Topics mit angegeben werden kann, wobei jedoch noch Duplikate und Übersetzungen berücksichtigt werden müssen. So ist es für die Bewertung nur von Bedeutung, an welcher Stelle der Ergebnisliste dieses Dokument auftaucht. Dabei bleibt allerdings unberücksichtigt, ob noch andere Dokumente in der Ergebnisliste für die Anfrage relevant sind. Es wurde vorgeschlagen, einen *Ad-hoc Task*, bei dem auch diese Dokumente in die Bewertung eingehen, einzuführen. Ein solcher kann aber nur durchgeführt werden, wenn Mittel für die Relevanzbewertung zur Verfügung gestellt werden (SIGURBJÖRNSSON ET AL. 2005b).

Aufgrund des *known-item* Ansatzes wird als primäres Evaluierungsmaß der Mean Reciprocal Rank (MRR) verwendet. Dieser gibt den Rang des relevanten Dokuments innerhalb der Ergebnisliste (Reciprocal Rank = 1 / Rang des ersten relevanten Dokuments in der Ergebnisliste) als Durchschnitt über die Anfragen an. Ein MRR von 1 würde demnach bedeuten, dass das relevante Dokument immer an erster Stelle zurückgegeben wurde, ein MRR von 0.25, dass es durchschnittlich an vierter Stelle liegt. Der Umfang der Ergebnisliste zu jeder Anfrage war bei WebCLEF auf 50 Treffer beschränkt. Ein weiteres Evaluierungsmaß ist die durchschnittliche Rate, mit dem das relevante Dokument in den ersten x Ergebnissen (Average Success @ 1, 5, 10, 20, 50) enthalten ist. Diese Maße erlauben Bewertungen in Hinblick auf eine hohe Präzision der Suche und sind im Kontext von Web Retrieval üblich, da angenommen wird, dass Nutzer im Internet häufig nur die ersten Ergebnisse einer Suche beachten (MISHNE & DE RIJKE 2006,504). Um die Bewertung und Postexperimente zu erleichtern wurden von den Veranstaltern Dateien mit den Ergebnislisten und ein Perl-Skript zur automatischen Überprüfung der Ergebnisse zur Verfügung gestellt.

3 Aktuelle Entwicklungen im Web Information Retrieval

Als Überblick zum aktuellen Forschungsstand im Bereich des Web Information Retrieval und des mehrsprachigen Information Retrieval sollen einige Systeme und Methoden aus den entsprechenden Tracks innerhalb der Evaluierungsinitiativen TREC[2] und CLEF[3] vorgestellt werden.

[2] http://trec.nist.gov/

3.1 TREC: Terabyte Track

Innerhalb der TREC-Initiative beschäftigt sich der Terabyte Track[4] mit der Suche in sehr großen Datensammlungen. Neben einem klassischen Ad-Hoc Task und einem Task in Hinblick auf die Recheneffizienz bei der Verarbeitung von Suchanfragen, gibt es in dieser Initiative auch einen den WebCLEF Tasks ähnlichen Task (Named Page Finding Task, vgl. Abschnitt 2: WebCLEF-Tasks und Bewertung). Die hierbei erfolgreichen Systeme setzen größtenteils auch webspezifische Retrieval Methoden ein, wie die Analyse der Linkstrukur (z.B. PageRank-Algorithmus[5] von Google), Berücksichtigung der HTML-Struktur der Dokumente oder der Texte von Links, die auf eine Seite zeigen (In-links).

So setzt beispielsweise das im Terabyte Track 2005 mit einem MRR von 0,441 zweitbeste System auf alle drei Varianten (CLARKE ET AL. 2005). Das von der der Universität Massachusetts erstellte System benutzt in seinem erfolgreichsten Run für den Named Page Finding Task eine Mischung aus verschiedenen Language Modelling-Techniken. Die Struktur der HTML-Dokumente wird analysiert und in Feldern für HTML-Title, Headings und Body indiziert. Ein weiteres Indexfeld enthält die Texte der Links, die auf das Dokument verweisen. Außerdem wird der PageRank-Algorithmus eingesetzt und die Zahl der In-Links berücksichtigt (METZLER ET AL. 2005).

Dass der Einsatz der genannten webspezifischen Methoden keine notwendige Voraussetzung für den Erfolg ist, zeigt das erstplazierte System der Tsinghua Universität, das hiervon nur die Indexierung der In-link Texte einsetzt und zusätzlich eine Analyse von Wortpaaren in den Anfragen vornimmt (ZHAO ET AL. 2005).
HAWKING & CRASWELL (2004) fassen für den Home Page Finding Task in TREC 2001 zusammen, dass der Einsatz von In-link Texten sehr effektiv ist, während die Berücksichtigung der Linkstruktur, wie die Zählung von In-links oder der PageRank-Algorithmus, weniger erfolgreich waren. Als vorteilhaft stellte sich die Analyse der URL in Hinblick auf die Position der Seite in der Hierarchie der Website heraus.

3.2 WebCLEF 2005: Mixed Monolingual Task

Die Systeme, die beim Mixed Monolingual Task von WebCLEF 2005 die besten Ergebnisse zeigten, waren das der Universität Glasgow und das der Firma Hummingbird.
Die Universität Glasgow setzte auf Feld-Spezifische Indexierung und sprachspezifisches Stemming (das Zurückführen der Terme auf ihre Stammformen). Allerdings konnte auch ohne Stemming oder mit Stemmingansätzen, die eigentlich für das Englische entwickelt wurden, vergleichbare, nur wenig schlechtere Ergebnisse erzielt werden. Das sprachspezifische Stemming konnte die Qualität der Ergebnisse sogar einschränken, wenn Fehler bei der Sprachidentifizierung auftraten. Während der Vorverarbeitung des Korpus wird bei diesem System die wahrscheinlichste Zeichenkodierung der Dokumente heuristisch auf der Basis von HTTP-Header und enthaltener Metadaten ermittelt und in UTF-8 kodiert (MACDONALD ET AL. 2005).

---

[3] http://www.clef-campaign.org/
[4] http://www-nlpir.nist.gov/projects/terabyte/
[5] vgl. Sergey Brin, Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Computer Networks and ISDN Systems 1998. http://www-db.stanford.edu/~backrub/google.html

Das System von Hummingbird benutzt nur das Title-Element aus der HTML-Struktur der Dokumente. Außerdem werden die URLs der Dokumente analysiert. Es wurde sowohl mit sprachunabhängigen Indexierungsmethoden als auch mit der Erstellung einzelner Indizes für elf Sprachen experimentiert. Eine höhere Gewichtung von Dokumenten, die höher in der Verzeichnishierarchie standen, hatte einen positiven Effekt für einen Teil der Topics (Homepage Finding, siehe Punkt 2: WebCLEF-Tasks und Bewertung). Auch sprachspezifische Stopwortlisten führten im Durchschnitt zu Verbesserungen. Wie auch für das System der Universität Glasgow war eine starke Gewichtung des HTML Title Elementes sehr effektiv (TOMLINSON 2005).

3.3 Multilinguale Verfahren im Kontext von WebCLEF

Für das sprachübergreifende Retrieval ist neben der Beherrschung verschiedener Sprachen auch die Übertragung zwischen den Sprachen von elementarer Bedeutung. Die bei WebCLEF 2005 eingesetzten Verfahren zur Übersetzung führten in diesem Kontext nicht zu Verbesserungen. Stattdessen wurden die besten Ergebnisse mit dem sprachunabhängigen Ansatz (Suche anhand der Anfragen und ihrer englischen Übersetzungen) der Universität Hildesheim erreicht. Auch die zweitplazierte Gruppe Miracle verwendete keine Übersetzungsverfahren. Hierbei zeigte sich auch, dass Anfragen, die Eigennamen enthielten, am erfolgreichsten behandelt werden konnten (SIGURBJÖRNSSON ET AL. 2005b). In welcher Form die in den anderen CLEF-Tracks entwickelten sprachspezifischen Verfahren für die komplexe Situation eines vielsprachigen Web-Korpus skalierbar sind, ist somit noch offen. Im Vordergrund der CLEF-Tracks stehen Verfahren aus den Bereichen der Übersetzung von Anfragen (hierbei beste Ergebnisse mit mehreren verschiedenen Ressourcen), der Anpassung von Indexierung und Anfrageverarbeitung an einzelne Sprachen und des Zusammenfassens der Ergebnisse aus den verschiedenen Sprachen (GONZALO & PETERS 2005).

4 Retrievalsystem der Universität Hildesheim

Die Universität Hildesheim hatte im Rahmen der CLEF-Initiative bereits seit mehreren Jahren Erfahrungen im multilingualen Retrieval gesammelt, so beispielsweise im Ad-hoc Track (HACKL et. al 2005). Vor diesem Hintergrund wurde 2005 ein System für den ersten Durchgang des WebCLEF-Tracks entwickelt, das mit seinem sprachunabhängigen Ansatz bei der multilingualen Suche das erfolgreichste im Feld der Teilnehmer war.

Für WebCLEF 2006 stand die Überarbeitung von grundlegenden Funktionen des Systems im Vordergrund, um Verbesserungen in beiden Bereichen, also auch für den Mixed Monolingual Track zu erreichen. Dabei wurde darauf Wert gelegt die Vorverarbeitung zu verbessern, um einen umfassenderen Index erstellen und mit mehreren Indexfeldern experimentieren zu können. Im Bereich des Retrieval wurde eine Blind Relevance Feedback Funktion zur Anfrageerweiterung implementiert. Die direkte Herangehensweise mit einem multilingualen Index wurde beibehalten, ebenso die multilinguale Suche ohne Übersetzung der Anfragen, die im letzten Durchgang zum Erfolg beim multilingualen Task geführt hatte. Obwohl dieser Task bei WebCLEF 2006 nicht stattfand, wurde zu Testzwecken auch mit dem neuen System ein multilingualer Run erstellt und eingereicht. Das entwickelte

System setzt als Basis-Suchmaschine die sehr effiziente Programmbibliothek Apache Lucene[6] ein und ist vollständig in der Programmiersprache Java implementiert.

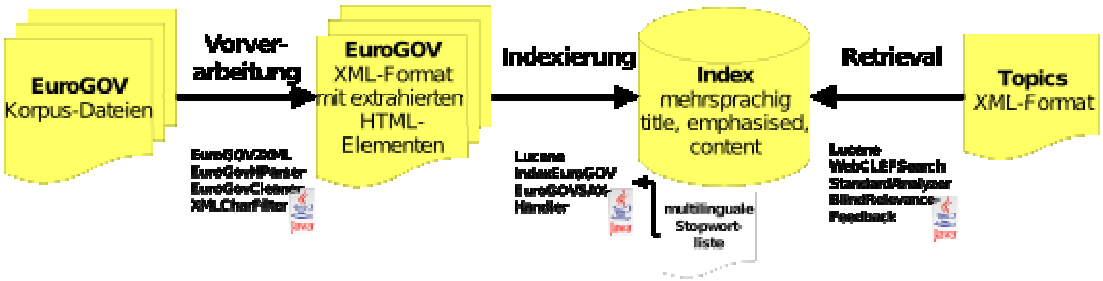

Abb. 2: System der Universität Hildesheim für WebCLEF 2006

4.1 Vorverarbeitung des EuroGOV-Korpus

Die Dateien, aus denen der Korpus besteht, werden zunächst in ein XML-konformes Format gebracht, damit bei der Indexierung mit einem XML-Parser darauf zugegriffen werden kann. Bei dem System zu WebCLEF 2005 waren nach der Vorverarbeitung nicht alle Dateien vollständig XML-konform. Dies führte dazu, dass einzelne Dokumente bei der Indexierung nicht erfasst werden konnten. Bei der Vorverarbeitung werden daher nun zunächst alle Zeichen gefiltert, die nicht XML-konform sind[7]. Solche Zeichen treten im Korpus häufig auf. Der Filter arbeitet direkt auf der Ebene des Zeichenstroms und damit sehr effizient. Eine vermutete Fehlerquelle sind die vielen unterschiedlichen Zeichenkodierungen, die im Korpus auftreten. Vor allem bei problematischen Kodierungen dürfte dies Auswirkungen auf die Retrievalqualität haben. Das Problem wird wohl nur durch die spezielle Behandlung jedes einzelnen Dokuments in Hinblick auf seine Kodierung behoben werden können. Nach der Filterung dieser Zeichen werden die Sonderzeichen in den URLs maskiert, und die in wohlgeformten XML-Dokumenten nicht zugelassenen verschachtelten CDATA-Elemente entfernt.

Um mittels des XML-Parsers auf Inhalte von einzelnen in den Dokumenten enthaltenen HTML-Elementen zugreifen zu können, werden diese aus dem CDATA-Bereich extrahiert und als neue Elemente in die XML-Struktur eingefügt. Berücksichtigt werden folgende Elemente: <title>, <h[1-6]>, , , <b> und <i>, da sie auf verschiedene Weise dazu dienen, Inhalte im HTML-Code inhaltlich oder optisch besonders hervorzuheben und diese daher potentiell bedeutungstragender als andere Inhalte sind. Im Laufe der Entwicklung wurde festgestellt, dass bei diesem Schritt wiederum Fehler im Dokument entstehen können, die daher in einem zweiten Arbeitsschritt beseitigt werden. Vor allem das mehrfache Ersetzen von Sonderzeichen durch XML-Entities (von & zu & zu & etc.), dessen Folgen häufig im Internet zu finden sind, führte zu unerwarteten Problemen. Aufgrund dieser Maßnahmen konnten alle Text- bzw. HTML-Dokumente indexiert werden.

[6] http://lucene.apache.org
[7] http://www.w3.org/TR/2004/REC-xml11-20040204/#charsets

4.2 Feldspezifische Indexierung

Für die Teilnahme 2006 sollte die Indexierungsstrategie durch die Berücksichtigung der HTML-Struktur der Dokumente erweitert werden. Die Inhalte wichtiger HTML-Elemente sollten in unterschiedlichen Feldern indiziert werden, um ihnen dann bei der Suche unterschiedliche Gewichtungen zuteilen zu können. Die Nutzung des HTML Titel-Elements hatte sich bereits für den Durchgang im Jahr 2005 als sehr effizient herausgestellt. Für die Implementierung war es eine wichtige Voraussetzung, dass die eingesetzte Suchmaschine Lucene die Erstellung von Indizes mit mehreren Feldern erlaubt. Aufgrund einer vorherigen Analyse der Häufigkeit, mit der die verschiedenen HTML-Elemente auftreten, wurde entschieden, die Inhalte von <title> und <h1>-Elementen zu einem Indexfeld ‚title', die Inhalte der anderen extrahierten Elemente in einem Feld ‚emphasised' zusammenzufassen, und dann beim Retrieval mit unterschiedlichen Gewichtungen der Indexfelder zu experimentieren. Der Inhalt der Dokumente wird einmal in einem Feld ‚content' komplett indiziert. Die Termvektoren, die für die anderen Felder berechnet und für das Blind Relevance Feedback benötigt werden, werden für dieses Volltext-Feld nicht berechnet, da der Schritt für den gesamten Dokumentinhalt zu rechenintensiv erschien. Stattdessen werden hierfür zusätzlich 50 Tokens aus dem Inhalt in einem eigenen Feld indiziert (content_cutoff). Der Text für dieses Feld wird aus der Mitte des Dokumentes genommen, da sich der bedeutungstragende Text einer Internetseite häufig nicht an ihrem Anfang befindet. Den Anfang einer Seite bilden häufig Menus und andere Navigationselemente, die in einem Webauftritt immer wieder in der gleichen Form auftreten und nicht spezifisch für ein einzelnes Dokument sind (CHEN ET AL. 2006). Aufwendigere Verfahren zum Aussortieren dieser Templates, wie beispielsweise durch den Vergleich verschiedener Seiten in Hinblick auf übereinstimmende Bestandteile, wurden nicht eingesetzt, um die Recheneffizienz des Systems nicht zu beeinträchtigen. Der Indexierungsvorgang auf einem Server mit 2 AMD Opteron 2,4 GHz Prozessoren und 8Gb Arbeitsspeicher dauerte 83 Stunden und der komplette Index mit Termvektoren beansprucht ungefähr 6Gb auf der Festplatte.

Vor der Indexierung werden an den Termen nur geringfügige Änderungen durchgeführt. Die Verarbeitung wird einem effizienten Standardmodul von Lucene überlassen, dem StandardAnalyzer. So wird außer der Entfernung von 's' -Endungen kein Stemming (das Zurückführen auf Wortstammformen) vorgenommen. Entsprechend dem Ansatz, mit einem multilingualen Index zu arbeiten, wurden keine sprachspezifischen Stemming-Algorithmen eingesetzt. Für WebCLEF 2005 hatte sich gezeigt, dass diese einfache Methode auch sprachunabhängig gute Ergebnisse liefern kann. Bestimmte Konstruktionen wie Akronyme oder E-Mail Adressen werden jedoch von der Analyzer-Klasse gesondert behandelt. Zusätzlich wurde eine Stopwortliste eingesetzt, um besonders häufig auftretende und damit wenig bedeutungstragende Wörter zu entfernen. Die bereits vorhandene multilinguale Stopwortliste, die 13 Sprachen umfasste, wurde um die im Korpus am häufigsten auftretenden Wörter erweitert. Diese erweiterte Liste beinhaltete 4722 Wörter. Die Idee, eine weitere titel-spezifische Stopwortliste einzusetzen, die z.B. auch automatisch erstellte und daher nicht bedeutungstragende Konstruktionen wie ‚no title' oder ähnliche Ausdrücke in anderen Sprachen entfernt hätte, wurde nicht verwirklicht. Eine Analyse der häufigsten Titel im EuroGOV-Korpus zeigte überraschenderweise, dass zwar einzelne Titel relativ häufig auftreten, diese aber trotzdem bedeutungstragend sind.

4.3 Gewichtung und Blind Relevance Feedback im Retrievalprozess

Beim Retrievalvorgang werden mit Hilfe des zuvor erstellten Indexes die besten Ergebnisse zu den Suchanfragen gesucht. Die Indexierung in mehrere Indexfelder schafft die Voraussetzung, um während des Retrievalvorganges mit unterschiedlichen Gewichtungen experimentieren zu können. Durch das Indexieren aller Dokumente mit Termvektoren konnte zusätzlich der Einsatz von Methoden des Blind Relevance Feedback (BRF) evaluiert werden. Termvektoren geben zu jedem Dokument die enthaltenen Terme und die Häufigkeit ihres Auftretens an. Für das BRF wird zunächst ein erster Suchdurchgang durchgeführt, um dann mit ausgewählten Termen aus den besten gefundenen Dokumenten (hierzu werden die Termvektoren benötigt) die ursprüngliche Suchanfrage zu erweitern und damit eine erneute Suche durchzuführen. Dahinter steht die Annahme, dass die im ersten Durchgang gefundenen besten Ergebnisse bereits der Anfrage entsprechen und weitere relevante Dokumente Ähnlichkeiten mit ihnen aufweisen sollten (GROSSMAN & FRIEDER 1998,84).

Die Experimente wurden durch die Tatsache vereinfacht, dass die Topics und die Ergebnisse aus dem Vorjahr als Testumgebung zur Verfügung standen und das von den Veranstaltern zur Verfügung gestellte Skript zur Überprüfung der Ergebnisse automatisiert zum Abschluss eines Suchvorgangs ausgeführt werden konnte, wodurch sofort Vergleichswerte zur Verfügung standen.

Um die Anfragen in das interne Format von Lucene zu übertragen, wird der Lucene QueryParser eingesetzt. Das Ranking der Ergebnisse basiert auf einer längennormalisierten tf-idf-Formel. Für eine Anfrage (Query) ‚q' und ein Dokument ‚d' wird der Rankingwert ‚score' wie folgt berechnet[8]:

$$\text{score}(q,d) = \sum_{t \text{ in } q}\left(\text{tf} \cdot \text{idf}^2 \cdot \{\text{Gewicht t in q}\} \cdot \{\text{Längennormalisierung}\}\right) \cdot \{\text{Überschneidung Anfrage/Dokument}\} \cdot \{\text{Query-Normalisierungsfaktor}\}$$

wobei:

- t = Term
- tf = term frequency (Anzahl des Auftretens des Terms im Dokument)
- idf = inversed document frequency: log(Anzahl Dokumente/(Anzahl Dokumente mit Term +1)) + 1
- {Gewicht t in q} = Gewichtung des Terms in der Anfrage
- {Längennormalisierung} = 1 / Wurzel(Anzahl Terme im Feld)
- {Überschneidung Anfrage/Dokument} = Faktor, der einbezieht, wie viele Terme der Anfrage auch im Dokument auftauchen: Terme, die in Anfrage und im Dokument enthalten sind / Anzahl der Worte in der Anfrage
- {Query-Normalisierungsfaktor}: hat keine Auswirkungen auf das Ranking, macht Rankingwerte verschiedener Anfragen vergleichbar

Gesucht wurde jeweils auf den Feldern ‚content' ‚emphasised' und ‚title'. Die Verwendung von ‚content_cutoff' statt des ‚content'-Feldes zeigte keine Vorteile, dieses Feld wurde daher nur für das BRF genutzt.

Damit die einzelnen Felder beliebig gewichtet werden konnten, wurde die Gewichtungsoption auch über die Kommandozeile verfügbar gemacht. Eine hohe Gewichtung des title-

[8] http://lucene.apache.org/java/docs/api/org/apache/lucene/search/Similarity.html

Feldes (20:0.1:1 im Verhältnis zu den beiden anderen Inhaltsfeldern, aber auch 10:5:1) brachte bei ersten Experimenten die größten Vorteile. Eine Suche nur auf dem Titelfeld führte im Vergleich zu schlechteren Resultaten. Die höhere Gewichtung der gemeinsam indexierten Inhalte von HTML-Überschriften (H2-H6) und anderer Elemente, die entweder der semantischen Hervorhebung oder der Hervorhebung im Schriftbild dienen (strong, em, bold und i), brachte hingegen nicht die gewünschten Ergebnisse. Eine differenzierte Indexierung der einzelnen Elemente könnte hier zu weiteren Erkenntnissen führen. Es könnten aber auch Verzerrungen in den Ergebnissen aufgetreten sein. Einmal, weil die Terme in den ‚title' und ‚emphasised' Feldern mehrmals (nämlich zusätzlich noch im Volltextfeld) indiziert wurden, was Auswirkungen auf die Gewichtung hat. Weiterhin sieht die verwendete Rankingformel von Lucene eine Längennormalisierung vor: Je kleiner der Inhalt eines Feldes ist, desto mehr Gewicht wird den einzelnen Worten gegeben. Dadurch, das die ‚titel' und ‚emphasised'-Felder kleiner sind als das ‚content' Feld, wird ihr relatives Gewicht bereits verstärkt. Entsprechende Versuche deuten im Nachhinein darauf hin, dass es effektiv ist, bei diesen Feldern mit sehr niedrigen Gewichtungen zu arbeiten, um die oben beschriebenen Effekte auszugleichen.

Vor dem Hintergrund des Mixed Monolingual Tasks werden zusätzlich die in den Metadaten der Topics angegebenen Internetdomains der gesuchten Seiten ausgewertet, um die Ergebnisse auf Dokumente in dieser Domain einzuschränken. Dies ist durch den Query-Filter von Lucene ohne großen Rechenaufwand möglich. Dieser hat zusätzlich den Vorteil, dass die Rankingergebnisse nicht weiter beeinflusst werden. Die Domains sind Teil der Dokument-IDs und werden während des Indexierens extrahiert und in ein eigenes Feld geschrieben. Ohne diesen Schritt musste eine Suche mit Platzhaltern auf den IDs ausgeführt werden, was zu Performance-Problemen führte. Dieser Domain-Filter wurde in allen Runs genutzt.

Ebenfalls zu Evaluierungszwecken können folgende Parameter von der Kommandozeile aus verändert werden: Die Anzahl der für das BRF eingesetzten Dokumente, die Anzahl der für die Anfrageerweiterung verwendeten Terme sowie die Gewichtung der Erweiterung relativ zur ursprünglichen Anfrage. Für das BRF wird eine an der Universität Hildesheim für den CLEF Ad-Hoc Track erstellte Implementierung eingesetzt, die auf den von Lucene bereitgestellten Termvektoren arbeitet und die Termgewichte über einen Robertson Selection Value berechnet (HACKL ET AL. 2005).

4.4 Vergleich zum System für WebCLEF 2005 und Ergebnisse aus den Experimenten

Zu den für WebCLEF 2006 implementierten Veränderungen am System gehören:

- die erschöpfendere Indizierung des Korpus (durch Volltextindizierung und Bereinigung von XML-Fehlerquellen für den Parser)
- die Feld-spezifische Indexierung weiterer HTML-Elemente (‚emphasised')
- eine Blind Relevance Feedback Funktion
- Einsatz von Metadaten (der Domain-Filter)

Insgesamt ist ein großer Teil der in den Experimenten festgestellten Verbesserungen (vgl. Abb. 3) auf den Einsatz des Domain-Filters zurückzuführen. Die Verbesserungen basieren

also auf dem Gebrauch von Metadaten und weniger auf den Verbesserungen am System. Um den Einfluss des Domain-Filters bereinigt, ergeben sich im Vergleich zu den Ergebnissen der letzten Teilnahme 2005 leicht verbesserte Werte für den MRR und deutliche Verbesserungen für den Average Success @ 50, also den durchschnittlichen Erfolg innerhalb der ersten 50 Ergebnisse. Zum Vergleich wurde einer der Runs auch ohne den Domain-Filter angegeben (Abb. 3: UhiTwoDF). Der Vergleich der beiden Runs ergibt, dass der Domain-Filter eindeutig zur Verbesserung des MRR beiträgt. Es kann angenommen werden, dass der Domain-Filter vor allem dazu beiträgt, die Genauigkeit der Suche zu erhöhen, indem die Position der gesuchten Ergebnisse in der Ergebnisliste gesteigert wird, da viele Dokumente, die nicht gültig sein können, ausgeschlossen werden. Die Tatsache, dass mit dem neuen System auch ohne Filter mehr gültige Dokumente gefunden werden (abzulesen an dem Erfolg nach 50 Ergebnissen), ist wahrscheinlich auf die erschöpfendere Indexierung zurückzuführen, da weder der Einsatz weiterer HTML-Elemente noch das Blind Relevance Feedback zu signifikanten Steigerungen führten.

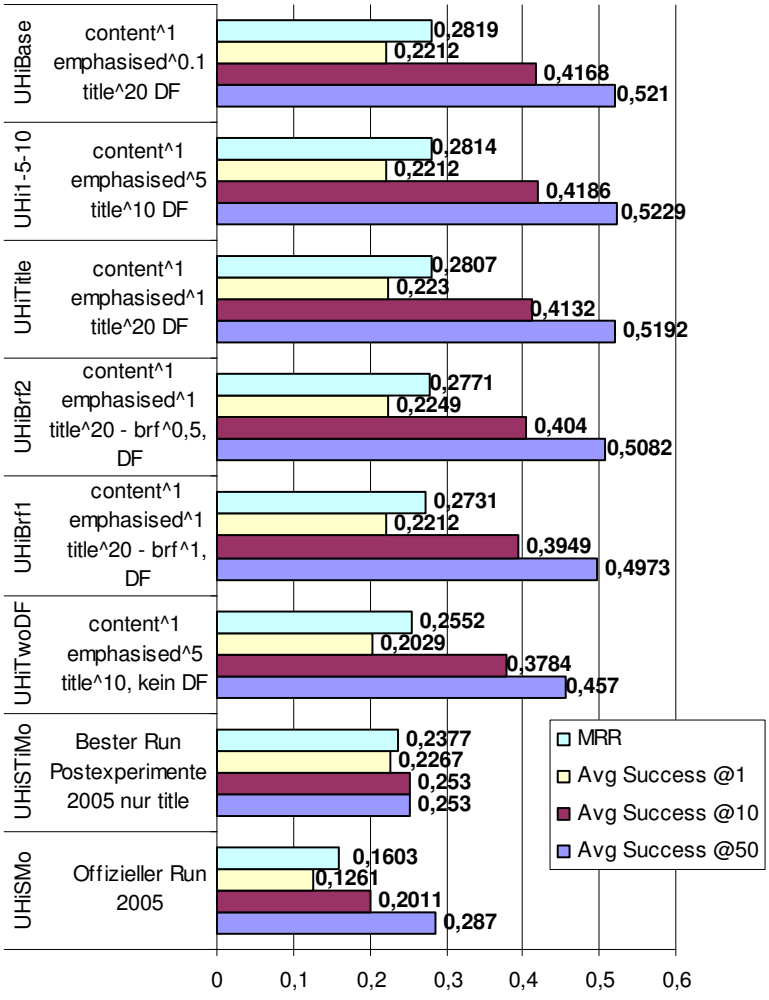

Abb. 3: Mixed Monolingual Task – WebCLEF 2005 und Postexperimente im Vergleich zu den diesjährigen Experimenten (DF=Domain-Filter)

Die Veränderungen am System führten auch im multilingualen Task zu Verbesserungen (vgl. Abb. 4). Hier zeigen sich die Verbesserungen bei der Indexierung, da keine weiteren Optimierungen für diesen Task durchgeführt wurden, und die Filterung nach Domains der Aufgabe entsprechend nicht in Frage kam. Besonders deutlich ist die Steigerung der insgesamt gefundenen Dokumente (Average Success @ 50).

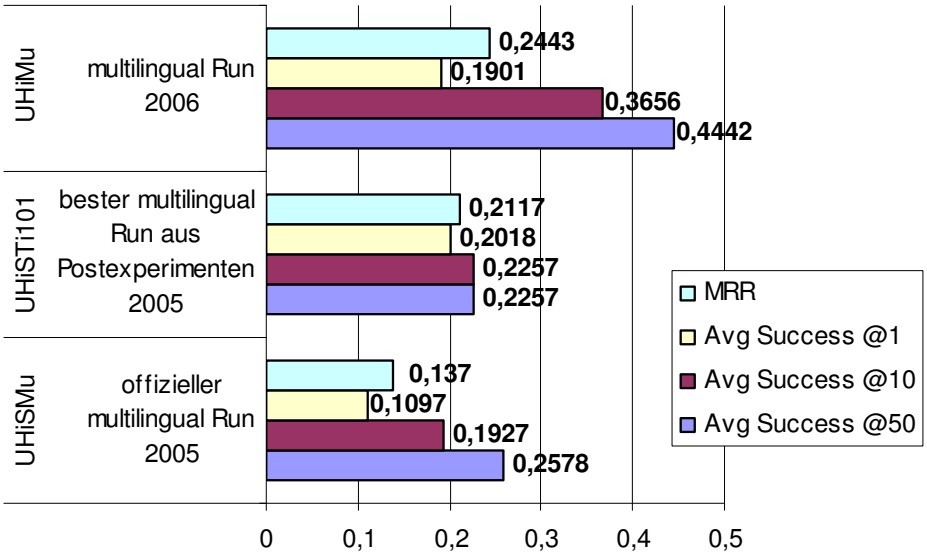

Abb. 4: Multilingual Task – WebCLEF 2005 und Postexperimente
im Vergleich zu den diesjährigen Experimenten

5 Ergebnisse WebCLEF 2006

5.1 Ergebnisse der Universität Hildesheim

Für die Teilnahme an WebCLEF 2006 wurden nach Experimenten mit den Topics des Vorjahres fünf Runs für den offiziellen Mixed Monolingual Task eingereicht sowie einer außer Konkurrenz für den multilingualen. Dabei wurde mit unterschiedlichen Gewichtungen der ‚emphasised' und ‚title' Felder sowie mit verschiedenen Parametern für BRF experimentiert (vgl. Tabelle 1). Bei allen Runs kam der Domainfilter zum Einsatz. Als Basis-Run wurde der beste Run aus den Experimenten gewählt.

Run	**Details**
UHiBase	Gewichte: content^1 emphasised^0.1 title^20, Domainfilter
UHiTitle	Gewichte: content^1 emphasised^1 title^20, Domainfilter
UHi1-5-10	Gewichte: content^1 emphasised^5 title^10, Domainfilter
UHiBrf1	Gewichte: content^1 emphasised^1 title^20 blind relevance feedback (Gewicht der Anfrageerweiterung: 1), Domainfilter
UHiBrf2	Gewicht: content^1 emphasised^1 title^20 blind relevance feedback (Gewicht der Anfrageerweiterung: 0.5), Domainfilter
UHiMu	(multilingual) content^1 emphasised^1 title^20 – Engl. Übersetzung^10

Tabelle 1: Parameter der eingereichten Runs der Universität Hildesheim für WebCLEF 2006

Beim Vergleich der Ergebnisse der einzelnen Runs fällt auf, dass diese relativ nah beieinander liegen (vgl. Tabelle 2). Die Variationen bei den Gewichtungen führten zu geringen Abweichungen. Dabei bestätigten sich die Erkenntnisse aus den Experimenten (vgl. Abschnitt 4: Vergleich zum System für WebCLEF 2005): Die starke Gewichtung des ‚title'-Feldes brachte Vorteile, die des ‚emphasised'-Feldes jedoch nicht. Dies reduziert die Wahrscheinlichkeit, dass die zusätzlich gewählten HTML-Elemente eine stärker diskriminierende Wirkung haben als sonstige Inhalte. Die Anfrageerweiterung mittels BRF brachte keine Vorteile, verschlechterte jedoch das Ergebnis nicht maßgeblich. Ob der Einsatz von BRF für die known-item Aufgabenstellung nützlich sein kann, ist somit noch offen. Die Verbesserungen gegenüber dem System aus WebCLEF 2005 lassen sich vor allem auf die Nutzung von Metadaten und die erschöpfendere Indexierung zurückführen.

Starke Unterschiede ergaben sich jedoch zwischen den unterschiedlichen Arten von Topics. So führten die automatisch erstellten Topics zu Ergebnissen, die weit unter denen der manuell erstellten Topics lagen (vgl. Tabelle 2). Dies kann daran liegen, dass die Topics teilweise problematisch waren (vgl. Abschnitt 2: Topics). Ein weiterer Grund dürfte aber auch die größere Sprachenvielfalt sein. Statt der elf Sprachen der manuellen Topics sind bei den automatischen erstellten potentiell alle Sprachen des Korpus enthalten, da bei der Erstellung alle 27 Domains genutzt wurden (BALOG ET AL. 2006a). Darauf weist auch der Unterschied innerhalb der manuell erstellten Topics. Die für 2006 neu hinzugekommen manuellen Topics, die weniger Sprachen (nur Niederländisch, Englisch, Deutsch, Ungarisch, Spanisch) umfassen als die Topics aus dem Vorjahr, führten zu besseren Ergebnissen. Die von der Art der Topics abhängigen Unterschiede in den Ergebnissen zeigten sich bei allen teilnehmenden Gruppen (vgl. Abb. 5).

	alle Topics		*automatisch erstellte Topics*		*manuell erstellte Topics*	
	MRR	Average success at 10	MRR	Average success at 10	MRR	Average success at 10
UHiBase	0,0795	0,1377	0,0346	0,0772	0,3076	0,4451
UHiTitle	0,0724	0,1253	0,0264	0,0630	0,3061	0,4420
UHi1-5-10	0,0718	0,1233	0,0242	0,0574	0,3134	0,4577
UHiBrf1	0,0677	0,1104	0,0220	0,0475	0,3000	0,4295
UHiBrf2	0,0676	0,1124	0,0221	0,0500	0,2989	0,4295
UHiMu	0,0489	0,0758	0,0083	0,0154	0,2553	0,3824

Tabelle 2: Ergebnisse Universität Hildesheim WebCLEF 2006 (ursprüngliche Topicauswahl), beste Ergebnisse grau hinterlegt

5.2 Überblick über die Teilnehmer

Aufgrund der Probleme mit den automatisch erstellten Runs wurde von den Veranstaltern im Nachhinein eine neue Topicauswahl erstellt, welche um alle Topics beschnitten wurde, in denen keine der Gruppen das korrekte Ergebnis liefern konnte. Die Ergebnisse wurden damit neu berechnet (Abb. 5). Aufgrund der zahlenmäßigen Dominanz der automatischen Topics und der mit diesen Topics verbundenen Probleme wurde zum Vergleich zusätzlich der Durchschnitt aus den beiden Ergebnissen für die automatischen und die manuellen Topics angegeben.

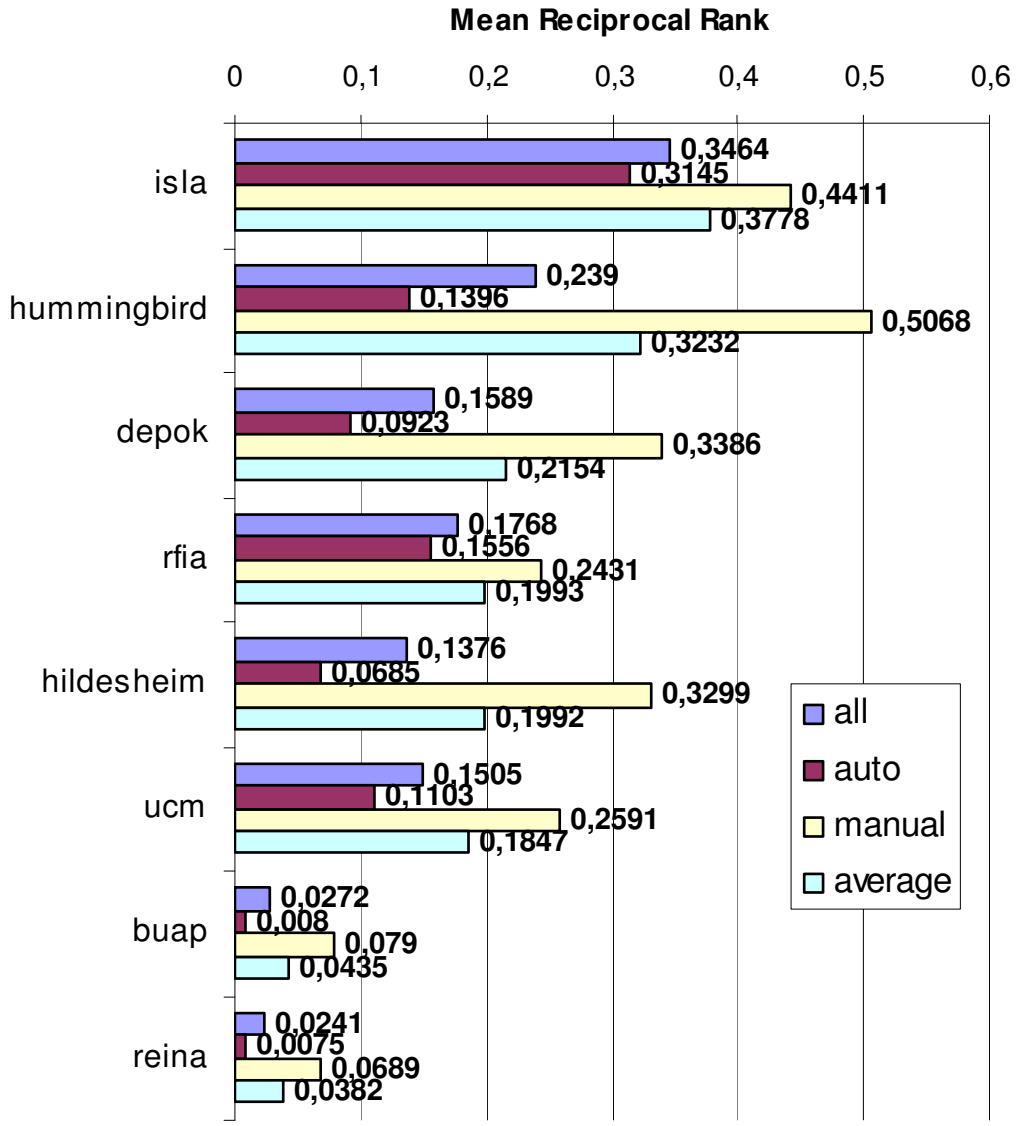

Abb. 5: Beste Runs der Teilnehmer in WebCLEF 2006 (MRR) mit der eingeschränkten Topicauswahl, *average* ist der Durchschnitt aus den Ergebnissen für die manuellen und automatischen Topics (BALOG ET AL. 2006)

Die Ergebnisse von zwei der acht teilnehmenden Gruppen setzen sich sowohl nach Durchschnittswert als auch nach absoluten Ergebnissen von den anderen ab (vgl. Abb. 5). Dies sind die Universität Amsterdam und die Firma Hummingbird. Die Ergebnisse von Hummingbird gehörten bereits bei WebCLEF 2005 zu den besten. Die vier folgenden Gruppen, u.a. die Universität Hildesheim, zeigen untereinander vergleichbare Ergebnisse, welche aber deutlich hinter denen der beiden ersten Gruppen zurückfallen. Der Teilnehmer mit den besten Ergebnissen im Mixed Monolingual Track des Vorjahres, die Universität Glasgow, nahm in diesem Jahr nicht teil.

Das System der Universität Amsterdam ist auf Basis von Lucene implementiert. Es wird kein Stemming vorgenommen. Die einzige sprachspezifische Maßnahme ist der Umgang mit griechischen und russischen Zeichenkodierungen. Es wurde ein Index für den kompletten Dokumentinhalt und einer für die ‚title'-Elemente erstellt und die Ergebnisse kombiniert, wobei mit verschiedenen Fusionsverfahren experimentiert wurde. (BALOG & DE RIJKE 2006)

Das System der Firma Hummingbird experimentierte mit der Gewichtung des HTML-Titel Feldes und der Analyse der URLs. Phrasen aus der Anfrage, die im Titel auftauchten, und Überschriften wurden höher gewichtet. Dies brachte positive Resultate, allerdings nur bei den manuell erstellten Topics. Stemming wurde nicht eingesetzt. (TOMLINSON 2006)

6 Zusammenfassung und Ausblick

Für die zweite Teilnahme am CLEF Web Track sollte das bestehende System durch die Indexierung mit verschiedenen Feldern unter Verwendung von HTML-Elementen verbessert werden. Die Verwendung des HTML-Titel Elementes erwies sich auch diesmal als vorteilhaft. Weitere Experimente mit der Feld-spezifischen Indexierung wären in Bezug auf die Gewichtung der Felder und dem Zusammenspiel mit der Rankingformel von Lucene sinnvoll.

Weiterhin konnte eine Blind Relevance Feedback Funktion implementiert werden. Die Verwendung von BRF hat allerdings nicht zu Verbesserungen der Retrievalergebnisse geführt. Es ist fraglich, ob die Anwendung im Kontext der known-item Suche sinnvoll ist. Allerdings bleibt auch hier Raum für Exprimente mit wichtigen Parametern, wie der Verwendung verschiedener Felder und der Gewichtung der Anfrageerweiterung unabhängig von der Gewichtung der ursprünglichen Anfrage.

Die Wirksamkeit des gewählten sprachunabhängige Indexierungsansatzes bestätigte sich sowohl durch die von uns durchgeführten Experimente, als auch dadurch, dass die bei WebCLEF 2006 besten Ergebnisse mit ganz oder weitgehend sprachunabhängigen Methoden erzielt wurden. Weiterhin konnten Erfahrungen mit der Implementierung eines Domain-Filters in Lucene gesammelt werden. Insgesamt zeigte das System der Universität Hildesheim geringfügige Verbesserungen im Vergleich zu 2005.

Zur weiteren Verbesserung des Systems sollte der Umgang mit den verschiedenen Zeichenkodierungen während der Vorverarbeitung verbessert werden. In zukünftigen Experimenten sollen außerdem qualitätsbasierte Bewertungsmethoden eingeführt werden, bei denen Informationen zum Dokumentlayout mit in die Bewertung der Dokumente eingehen (MANDL 2006). Auch hierfür ist eine verbesserte Vorverarbeitung des EuroGOV-Korpus notwendig, um während der Indexierung einen direkten Zugriff auf die HTML-Struktur der Dokumente zu ermöglichen.

Literaturverzeichnis

Balog, Krisztian; Azzopardi, Leif; Kamps, Jaap; de Rijke, Maarten (2006): Overview of WebCLEF 2006. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/balogOCLEF2006.pdf Verifiziert 15.9.2006

Balog, Krisztian; de Rijke, Maarten (2006): The University of Amsterdam at WebCLEF 2006 http://www.clef-campaign.org/2006/working_notes/workingnotes2006/balogCLEF2006.pdf Verifiziert 15.9.2006

Chen, L.; Ye, S.; Li, X. (2006): Template Detection for Large Scale Search Engines. In: Proceedings ACM Symposium on Applied Computing ACM Press. S. 1094-1098.

Clarke, Charles L. A.; Scholer, Falk, Soboro, Ian (2005): The TREC 2005 Terabyte Track. http://trec.nist.gov/pubs/trec14/papers/TERABYTE.OVERVIEW.pdf Verifiziert 15.9.2006

Gonzalo, Julio; Peters, Carol (2005): The Impact of Evaluation on Multilingual Text Retrieval. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, New York 2005 S. 603-604

Grossman, David A.; Frieder, Ophir (1998): Information retrieval : algorithms and heuristics. Kluwer, Boston

Hackl, René; Mandl, Thomas; Womser-Hacker, Christa (2005): Ad-hoc Mono- and Multilingual Retrieval Experiments at the University of Hildesheim. In: Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005. Sept. 2005, Wien.

Hawking, David; Craswell, Nick (2004): Very Large Scale Retrieval and Web Search (Preprint version). http://es.csiro.au/pubs/trecbook_for_website.pdf Verifiziert 15.9.2006

Heuwing, Ben; Mandl, Thomas; Strötgen, Robert (2006): Multilingual Web Retrieval Experiments with Field Specific Indexing Strategies for CLEF 2006 at the University of Hildesheim. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/heuwingCLEF2006.pdf Verifiziert 19.9.2006

Jensen, Niels (2005a) Web Information Retrieval am Beispiel des WEB-GOV Korpus. Magisterarbeit Internationales Informationsmangagement, Universität Hildesheim.

Jensen, Niels (2005b) Mehrsprachiges Information Retrieval mit einem WEB-Korpus. In: Mandl, Thomas; Womser-Hacker, Christa (Hrsg.) (2006): Effektive Information Retrieval Verfahren in Theorie und Praxis: Ausgewählte und erweiterte Beiträge des Vierten Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2005) Hildesheim, 20.7.2005. Konstanz: Universitätsverlag [Schriften zur Informationswissenschaft 45] S. 235-244.

Jensen, Niels; Hackl, René; Mandl, Thomas; Strötgen, Robert (2006): Web Retrieval Experiments with the EuroGOV Corpus at the University of Hildesheim. In: Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, Revised Selected Papers. Springer [LNCS 4022]

Macdonald, Craig; Plachouras , Vassilis; He, Ben; Lioma, Christina; Ounis, Iadh (2005): University of Glasgow at WebCLEF 2005: Experiments in per-field normalisation and language specific stemming. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/macdonald05.pdf Verifiziert 15.9.2006

Mandl, Thomas (2006): Implementation and Evaluation of a Quality Based Search Engine. In: Proceedings of the 17th ACM Conference on Hypertext and Hypermedia (HT '06) Odense, Denmark, 22.–25 August. ACM Press.

Metzler, Donald; Strohman, Trevor; Zhou, Yun; Croft, W.B. (2005): Indri at TREC 2005: Terabyte Track. http://trec.nist.gov/pubs/trec14/papers/umass-tera.pdf Verifiziert am 15.9.2006

Mishne, Gilda; de Rijke, Maarten (2005): Boosting Web Retrieval Through Query Operations. In: Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 2005 Proceedings. Springer [LCNS 3408] S. 502-516

Santos, Diana; de Rijke, Maarten (2005): WebCLEF 2005 workshop report. http://ilps.science.uva.nl/WebCLEF/WebCLEF2005/Report/index.html Verifiziert 8.9.2006

Sigurbjörnsson, Börkur; Kamps, Jaap; de Rijke, Maarten (2005a): Blueprint of a Cross-Lingual Web Retrieval Collection. In: Journal of Digital Information Management, vol. 3 (1) S. 9-13.

Sigurbjörnsson, Börkur; Kamps, Jaap; de Rijke, Maarten (2005b): Overview of WebCLEF 2005. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/sigurbjornsson05.pdf Verifiziert 15.9.2006

Tomlinson, Stephen (2005): European Web Retrieval Experiments with Hummingbird SearchServerTM at CLEF 2005. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/tomlinson05.pdf Verifiziert 15.9.2006

Tomlinson, Stephen (2006): European Web Retrieval Experiments at WebCLEF 2006 http://www.clef-campaign.org/2006/working_notes/workingnotes2006/tomlinsonWebCLEF2006.pdf Verifiziert 15.9.2006

Zhao, Le; Ceng, Rongwei; Ma, Shaoping; Jin, Yijiang; Zhang, Min (2005): THUIR at TREC 2005 Terabyte Track. http://trec.nist.gov/pubs/trec14/papers/tsinghuau-ma.tera.pdf Verifiziert 15.9.2006

Benutzerorientierte Bewertungsmaßstäbe für Information Retrieval Systeme: Der *Robust Task* bei CLEF 2006

Thomas Mandl

Universität Hildesheim
Informationswissenschaft
Marienburger Platz 22
31141 Hildesheim
mandl@uni-hildesheim.de

Zusammenfassung

Die Qualität von Antworten im Information Retrieval schwankt zwischen einzelnen Anfragen sehr stark. Die Evaluierung im Information Retrieval zielt in der Regel auf eine Optimierung der durschnittlichen Retrieval-Qualität über mehrere Testanfragen (Topics). Sehr schlecht beantwortete Anfragen wirken sich besonders negativ auf die Zufriedenheit des Benutzers aus. Neue Ansätze zur Evaluierung der Robustheit von Systemen werten daher die schwierigen Anfragen stärker. Im Rahmen des Cross Language Evaluation Forum (CLEF) wurde 2006 ein *Robust Task* durchgeführt. Der Artikel zeigt die Gründe für Entwicklung dieser Aufgabenstellung nach, referiert die Ergebnisse und verweist auf zukünftige Planungen.

Abstract:

The quality of the responses of information retrieval system differs greatly between queries. The evaluation in information retrieval usually considers the average retrieval quality over several queries. However, queries which lead to poor results have a higher impact on the satisfaction of the user. New appraoches try to consider the users perspective by emphazising the hard topics. Within the Cross Language Evaluation Forum (CLEF), a robust task has been established in 2006. This article reports the reasons for the development, the task design and shows results and potenntial future trends.

1 Einleitung

Die Evaluierung im Information Retrieval zielt in der Regel auf eine Optimierung der durschnittlichen Retrieval-Qualität über mehrere Testanfragen (Topics). Der Eindruck des Benutzers von einem System wird aber stark durch sehr schlecht beantwortete Anfragen geprägt. Wird eine Nullantwort durch eine geringfügige Verbesserung für ein Topic vermieden, so hilft dies dem Benutzer meist mehr als eine geringfügige Steigerung bei einer ohnehin gut beantworteten Anfrage. Die *Mean Average Precision* (MAP) als wichtigstes Maß bei allen Evaluierungen weist beiden Fällen die gleiche Bedeutung zu.

Neuere Entwicklungen in der Evaluierungsforschung versuchen, diese Benutzerperspektive in das Zentrum zu rücken. Vor diesem Hintergrund ist die Debatte um adäquate Evaluierungsmaße im Information Retrieval wieder aufgegriffen worden. Zwar ist bekannt, dass die meisten Evaluierungsmaße eine starke Korrelation untereinander aufweisen (BUCKLEY & VOORHEES 2005). Trotzdem ist die Analyse benutzerorientierter Maßzahlen sinnvoll.

Ein wichtiges Maß, welches häufig eingesetzt wird und das die schwierigeren Anfragen stärker gewichtet, ist der geometrische Durchschnitt oder der geometrische Mittelwert. Er berechnet sich als die n-te Wurzel aus dem Produkt der zu mittelnden Einzelwerte.

$$geoAve = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Die Einzelwerte stellen im Information Retrieval die Ergebnisse der einzelnen Aufgaben (Topics, Anfragen) dar. Denkbar wäre auch, die Ergebnisse der einzelnen Topics bereits anders zu gewichten. Viele Benutzer, vor allem bei Internet-Suchmaschinen bewerten die Precision unter den ersten Treffern besonders hoch und legen geringeren Wert auf den Recall.

Für die Bewertung der Robustheit hat sich besonders im Rahmen der Text Retrieval Conference (TREC) das geometrische Mittel etabliert. Abbildung 1 zeigt ein konstruiertes Beispiel, welches dessen Wirkung verdeutlicht.

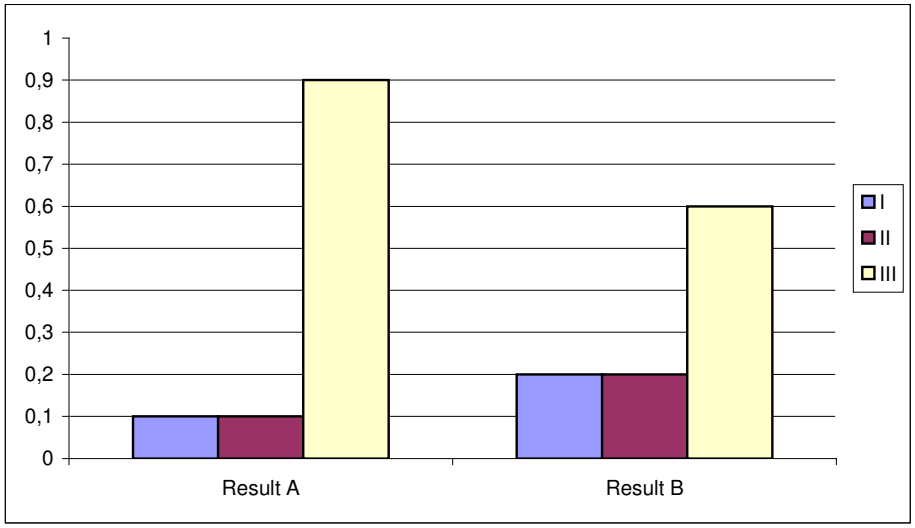

Abb. 1: Exemplarische Ergebnisse zweier Retrieval Systeme

Tabelle 1: Auswertung der Ergebnisse aus Abbildung 1

Topic	System	Ergebnis	Topic	System	Ergebnis
1	A	0,1	1	B	0,2
2	A	0,1	2	B	0,2
3	A	0,9	3	B	0,6
GeoAve	A	0,21	GeoAve	B	**0,29**
MAP	A	**0,37**	MAP	B	0,33

Abbildung 1 und Tabelle 1 gegen die Ergebnisse von zwei Systemen für drei Anfragen (Topics) wider. System A erreicht bei Topic 3 ein optimales Ergebnis und bei Topic 1 und 2 nur ein schlechtes Ergebnis. Während das System A bei der Bewertung durch die durchschnittliche Precision besser abschneidet, weist das geometrische Mittel System B die bessere Gesamtbewertung zu. Da der Benutzer in den meisten Situationen die Verbesserung von System B für Topic 1 und 2 höher bewerten wird, als die Verschlechterung auf hohem Niveau bei Topic 3, entspricht das geometrische Mittel eher der Benutzer-Perspektive.

Es zeigt sich in vielen Evaluierungen, dass Ergebnisse wie das oben konstruierte durchaus vorkommen (siehe Abschnitt 3).

2 Robustheit als Ziel im Information Retrieval

Die Variabilität zwischen den Topics ist bei allen Evaluierungen meist größer als die zwischen den Systemen (BUCKELY 2004, BRASCHLER 2003). Dementsprechend verspricht die Analyse der einzelnen Topics großes Potential für die Verbesserung der Retrieval-Ergebnisse (MANDL & WOMSER-HACKER 2003, HARMAN & BUCKLEY 2004):

> „More work needs to be done on customizing methods for each topic“ (HARMAN 2005)

Für die Steigerung der Robustheit ist es erforderlich, besonders die Qualität der Systeme für die schwierigen Topics zu erhöhen. Dazu ist sowohl die automatische Vorhersage der Schwierigkeit als auch die Analyse und besondere Behandlung der vermutlich schwierigen Topics nötig.

Der Workshop *Reliable Information Access* (RIA)[1] versuchte das zweite Problem zu bearbeiten (HARMAN 2004). RIA wurde vom das *National Institute for Standards and Technology* (NIST) in Gaithersburg, Maryland, USA veranstaltet und dort durchgeführt. Führende Forschungsgruppen trafen am NIST zusammen und arbeiteten dort mit ihren Systemen vor Ort an den besonders schwierigen Topics aus der TREC Evaluierungsinitiative.

Die Text Retrieval Conference (TREC[2]) findet seit 1992 statt und hat eine neue Ära in der Evaluierung im Information Retrieval eingeläutet. Im Jahr 2004 veranstaltete das *National Institute for Standards and Technology* (NIST) bereits den dreizehnten Workshop dieser ersten großen Evaluierungsinitiative (VOORHEES & BUCKLAND 2004). Der ad-hoc Track war zu Beginn von TREC der wichtigste und bedeutendste Track (WOMSER-HACKER 2004) und fand bis 1999 statt. Danach wurde ad-hoc Retrieval in anderen Tracks forgeführt. Eine Variante ist der *Robust Retrieval Track*, der 2003, 2004 und 2005 durchgeführt wurde. Die Datengrundlage für den Robust Track bei TREC 2003 bestand aus ca. 500.000 Dokumenten aus alten TREC ad-hoc Tracks. Die Bewertungsmethodologie wurde modifiziert und die Gruppen wurden informiert, dass die Bewertung hauptsächlich anhand des geometri-

[1] http://ir.nist.gov/ria/
[2] http://trec.nist.gov

schen Mittelwerts erfolgen würde. Dadurch richteten die Organisatoren des *Robust Retrieval Track* die Evaluierungsmethodik stärker an den Bedürfnissen des Benutzers aus. Der Fokus lag nicht auf der durchschnittlichen Performanz der Systeme über alle Topics, sondern auf einer stabilen Performanz über alle Topics. Dazu mussten Ergebnisse für schwierige Topics stärker werten als einzelne herausragende Ergebnisse für eher leichte Topics, was durch die geometrische Mittelwertbildung erreicht wurde (VOORHEES 2005a, VOORHEES 2005b). Eine weitere Aufgabe im *Robust Retrieval Track* bestand darin, die Topics nach ihrer Schwierigkeit zu ordnen.

Der RIA Workshop widmete sich der Analyse der Gründe für das Scheitern der Systeme bei bestimmten Topics. Aus dem RIA Workshop entstand der *Robust Track* im Rahmen von TREC.

Ein weiterer kürzlich durchgeführter Workshop widmete sich dem Thema der Schwierigkeit einzelner Topics. Der Workshop *Predicting Query Difficulty - Methods and Applications* bei SIGIR 2005[3] versuchte, den Schwierigkeitsgrad einzeler Anfragen vorherzusagen.

3 Robust Task bei CLEF 2006

Basierend auf den Erfahrungen bei TREC wurde auch im Rahmen von CLEF ein *Robust Task* durchgeführt, der vom Autor organisiert wurde.

Die CLEF-Initiative[4] (MANDL 2005, PETERS ET AL. 2005) etablierte sich im Jahr 2000. Seitdem steigt die Zahl der Teilnehmer bei CLEF stetig an und die Forschung zu mehrsprachigen Information Retrieval Systemen gewinnt an Dynamik. CLEF folgt dem Modell von TREC und schuf eine mehrsprachige Kollektion mit Zeitungstexten. Inzwischen umfasst die Dokument-Kollektion für das ad-hoc Retrieval die Sprachen Englisch, Französisch, Spanisch, Italienisch, Deutsch, Holländisch, Schwedisch, Finnisch, Portugiesisch, Bulgarisch, Ungarisch und Russisch. Mehrere weitere Tracks wie *Question Answering*, *Web-Retrieval*, *Spoken Dokument Retrieval* oder *Geographic CLEF* untersuchen bestimmte Aspekte des mehrsprachigen Retrieval.

Vor der Einführung eines *Robust Tasks* mussten hierfür die Ergebnisse früherer CLEF Experimente analysiert werden. Die Einführung neuer Evaluierungsmaße lohnt nur, wenn die Korrelation zwischen den traditionellen und den neuer Maßen nicht zu hoch ist. Dazu wurden die Ergebnisse von CLEF 2001 bis 2003 untersucht. Um die Ähnlichkeit der Rankings auf der Basis des MAP und des geometrischen Mittelwerts zu bestimmen, wurde der Spearman Rang-Korrelations-Koeffizient berechnet. Für eine identische Rangfolge liefert der Koeefizient den Wert 1, für eine umgekehrte Rangfolge den Wert –1. Die folgende Tabelle 3 zeigt die Korrelationen für einige ausgewählte Experiment-Typen.

[3] http://www.haifa.ibm.com/sigir05-qp/
[4] http://www.clef-campaign.org

Tabelle 2: Rang Korrelation für CLEF Experimente nach Spearman

Task	**Topic Sprache**	**CLEF Jahr**	**Korrelation**
Mono	German	2001	0,91
Multi	English	2001	0,96
Mono	Spanish	2001	0,93
Bi	English	2002	0,98

Die Korrelationen sind sehr hoch, jedoch sind die Rangfolgen nicht identisch. Bei genauerer Analyse erweisen sich auch relevante Unterschiede. Die Top-Systeme unterscheiden sich in mehreren Fällen. Um dies zu illustrieren, zeigt Tabelle 3 die Position des besten Systems in der MAP Rangfolge auf einer nach dem geometrischen Mittel sortierten Skala.

Tabelle 3: Rang der besten Systeme nach MAP im geoAve Ranking

Task	**Topic Sprache**	**CLEF Jahr**	**Rang**
Mono	German	2001	2
Multi	English	2001	1
Mono	Spanish	2001	1
Bi	English	2002	10

Auch in anderen Fällen ändern sich Positionen dramatisch. Zum Beispiel fällt das zweitbeste System für CLEF 2001 (Topic Sprache Englisch) auf Platz 32 von 56 Teilnehmern.

Bisher wurde als Definition für die Schwierigkeit eines Topics ein niedriger maximaler MAP-Wert genutzt. Dies entspricht dem Vorgehen der meisten Forscher (EGUCHI ET AL. 2002, KWOK 2005, CRONEN-TOWNSEND ET AL. 2002, MOTHE & TANGUY 2005). Die genaue Analyse der CLEF-Ergebnisse weckt aber Zweifel an dieser Perspektive und der Definition. Zum einen kann sowohl das geometrische Mittel aus den Systemen benutzt werden. Dann wird der Einfluss besonders schlechter Systeme erhöht. So könnte eine Menge von Topics identifiziert werden, die sich für viele System als besonders schwierig erweisen. Daneben könnte auch das Ergebnis des besten Systems für dieses Topic genutzt werden. Dies kann als Wert für das Maximum gelten, welches für dieses Topic erzielt werden kann. Die folgende Tabelle 4 zeigt die Größe der Schnittmenge zwischen den beiden alternativen Definitionen und den zehn schwierigsten Topics nach dem durchschnittlichen MAP Wert aller Systeme.

Tabelle 4: Anzahl unterschiedlicher schwieriger Topics

Task	Mono	Multi	Mono	Bi
Topic Sprache	Deutsch	Englisch	Spanisch	Englisch
CLEF Jahr	2001	2001	2001	2002
Geometr. Mittel	2	2	2	2
Bestes System	3	3	2	2

Es erweist sich, dass die Schnittmenge zwischen den Definitionen sehr klein ist. Weitere Definitionen scheinen denkkbar. Die Menge der relevanten Dokumente in der Kollektion mag ein für den Benutzer sehr einleuchtendes Maß darstellen. Interessant sollte neben der absoluten Definition vor allem das Verbesserungspotential sein, wofür die Varianz zwischen den Systemen berücksichtigen werden müsste. Die Frage, welche Topics eigentlich schwer sind, bleibt letztlich offen.

Das Task Design für den *Robust Task* versuchte vor allem einen große Menge von Dokumenten und Topics zu finden, für die ohne weitere Relevanz-Urteile eine Analyse durchgeführt werden konnte. Die Wahl fiel auf die CLEF Jahhre 2001, 2002 und 2003, in denen für eine Reihe von Kernsprachen eine weitegehend konstante Dokumentenkollektion benutzt wurde. Somit standen aus diesen drei Jahren 160 Topics zur Verfügung.

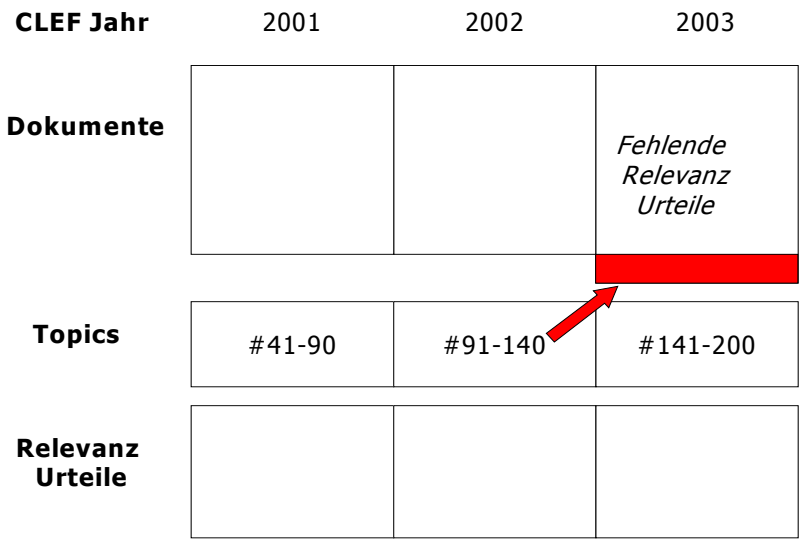

Abbildung 2: Dokumente und Topics für den *Robust Task*

Die folgende Tabelle 5 listet die verwendeten Kollektionen auf. Insgesamt umfasst die sechssprachige Datensammlung 1,35 Millionen Dokumente mit 3,6 Gigabyte Text.

Tabelle 5: Korpora für den *Robust Task*

Sprache	**Kollektion**
Englisch	LA Times 94, Glasgow Herald 95
Französisch	ATS (SDA) 94/95, Le Monde 94
Italienisch	La Stampa 94, AGZ (SDA) 94/95
Holländisch	NRC Handelsblad 94/95, Algemeen Dagblad 94/95
Deutsch	Frankfurter Rundschau 94/95, Spiegel 94/95, SDA 94
Spanisch	EFE 94/95

Die in Abbildung 2 angedeutete Inkonsistenz zwischen Relevanz-Bewertungen und Kollektion entstand durch geringfügige Änderungen der deutschen Kollektion. Die Teilnehmer

sollten das Wissen darüber, dass für einige Topics in der neueren Kollektion keine bewerteten Dokumente enthalten sind, nicht ausnutzen. Dies konnte zu Schwierigkeiten bei der Optimierung der Systeme führen und wurde von einigen Teilnehmern kritisiert. Das Eingrenzen der Ergebnisse für diese Anfragen auf die bewerteten Korpora hätte zu einer Verbesserung von bis zu 10% bei der MAP führen können (SAVOY 2006).

Die Identifikation von besonders schwierigen Topics hatte sich als problematisch erwiesen. Die Schwierigkeit war zwischen den Sprachen sehr unterschiedlich und die Analse bestätigte letztlich ein Ergebnis das Robust Tracks bei TREC. Eine Anfrage ist nicht *per se* schwierig, sondern nur in Zusammenspiel mit einer Kollektion (VOORHEES 2005b).

Somit wurde für den Robust Task bei CLEF keine Menge von schwierigen Topics definiert, sondern die Topics wurden zufällig in zwei Mengen geteilt. Während 60 Topics zum Training dienten, sollten die übrigen 100 als Testdaten benutzt werden.

4 Ergebnisse

Insgesamt betiligten sich acht Gruppen an dem Robust Task und reichten 133 Runs (Experimente) ein (DI NUNZIO ET AL. 2006).

Table 6 : Teilnehmer am CLEF Robust Task 2006 (DI NUNZIO ET AL. 2006)

U. Coruna & U. Sunderland (Spanien & UK)	Hummingbird Core Tech. (Kanada)
U. Jaen (Spanien)	U. Neuchatel (Schweiz)
DAEDALUS & Madrid Univs. (Spanien)	Dublin City U. – Computing (Irland)
U. Salamanca – REINA (Spanien)	U. Hildesheim – Inf. Sci. (Deutschland)

Tabelle 7: Anzahl der eingereichten Ergebnisse für den CLEF *Robust Task* 2006

Task	**Sprache**	**Anzahl Test Runs**	**Anzahl Training Runs**	**Anzahl Gruppen**
mono	en	13	7	6
	fr	18	10	7
	nl	7	3	3
	de	7	3	3
	es	11	5	5
	it	11	5	5
bi	it->es	8	2	3
	fr->nl	4	0	1
	en->de	5	1	2
multi	multi	10	3	4

Die folgenden Tabellen zeigen die Ergebnisse der besten Systeme für mono-linguale Experimente (aus DI NUNZIO ET AL. 2006)

Tabelle 8: Ergebnisse für mono-linguale Experimente beim *Robust Task* CLEF 2006

Track	Teilnehmer Rang					
	1	2	3	4	5	Differenz
Deutsch	hummingbird	colesir	daedalus			1. vs 3.
MAP	48,30%	37,21%	34,06%			41,81%
GMAP	22,53%	14,80%	10,61%			112,35%
Run	humDE06Rtde	CoLesIRdeTst	deFSdeR2S			
Italienisch	hummingbird	reina	dcu	daedalus	colesir	1. vs 5.
MAP	41,94%	38,45%	37,73%	35,11%	32,23%	30,13%
GMAP	11,47%	10,55%	9,19%	10,50%	8,23%	39,37%
Run	humIT06Rtde	reinaITtdtest	dcudescit1005	itFSitR2S	CoLesIRitTst	
Spanisch	hummingbird	reina	dcu	daedalus	colesir	1. vs 5.
MAP	45,66%	44,01%	42,14%	40,40%	40,17%	13,67%
GMAP	23,61%	22,65%	21,32%	19,64%	18,84%	25,32%
Run	humES06Rtde	reinaEStdtest	dcudescsp12075	esFSesR2S	CoLesIResTst	

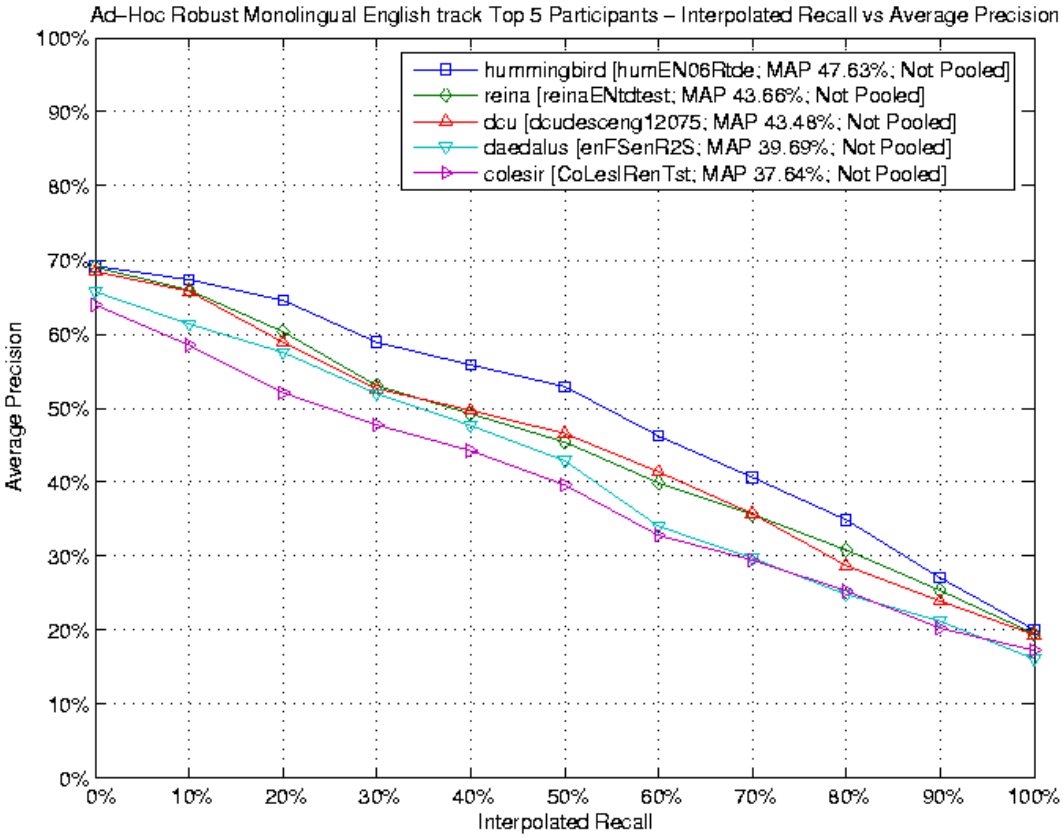

Abb. 3: Ergebnisse des Robust Task bei CLEF 2006 für Englisch (aus DI NUNZIO ET AL. 2006)

Tabelle 9: Ergebnisse für mono-linguale Experimente beim *Robust Task* CLEF 2006

Track	Teilnehmer Rang					
	1	2	3	4	5	Differenz
Holländisch	hummingbird	daedalus	colesir			1. vs 3.
MAP	51,06%	42,39%	41,60%			22,74%
GMAP	25,76%	17,57%	16,40%			57,13%
Run	humNL06Rtde	nlFSnlR2S	CoLesIRnlTst			
Englisch	hummingbird	reina	dcu	daedalus	colesir	1. vs 5.
MAP	47,63%	43,66%	43,48%	39,69%	37,64%	26,54%
GMAP	11,69%	10,53%	10,11%	8,93%	8,41%	39,00%
Run	humEN06Rtde	reinaENtdtest	dcudesceng 12075	enFSenR2S	CoLesIRenTst	
Französisch	unine	hummingbird	reina	dcu	colesir	1. vs 5.
MAP	47,57%	45,43%	44,58%	41,08%	39,51%	20,40%
GMAP	15,02%	14,90%	14,32%	12,00%	11,91%	26,11%
Run	UniNEfrr1	humFR06Rtde	reinaFRtdtest	dcudescfr12075	CoLesIRfrTst	

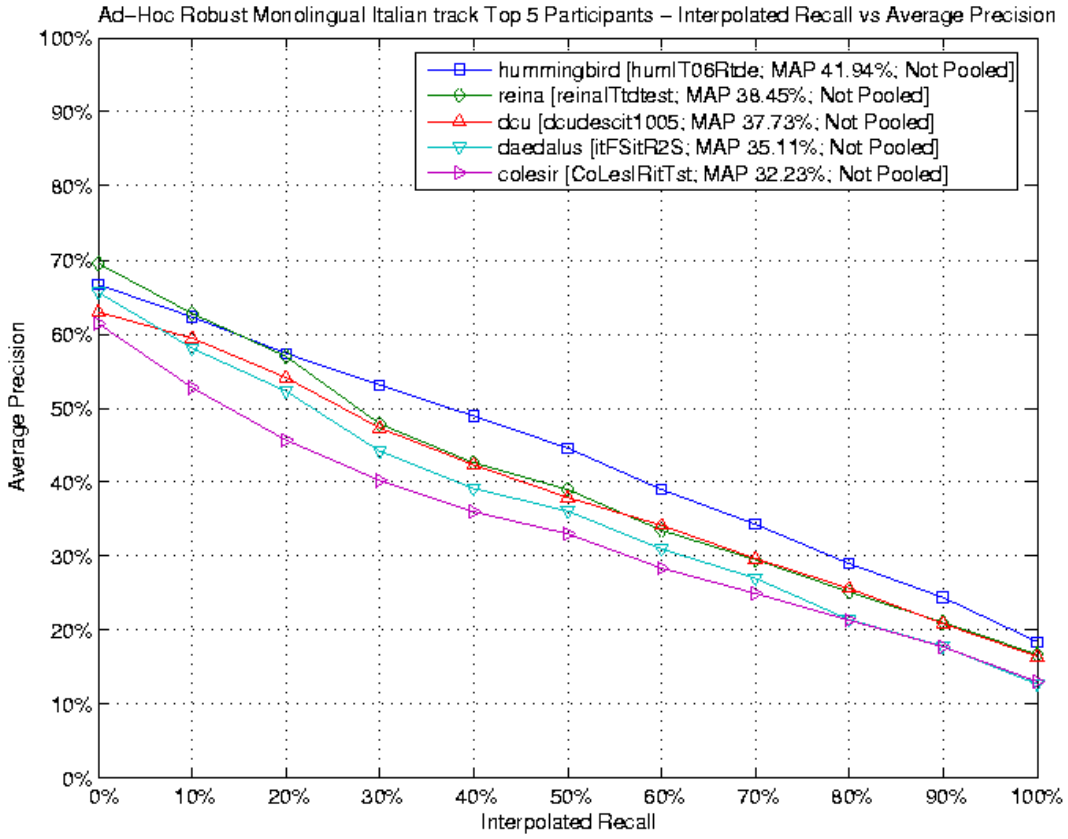

Abb. 4: Ergebnisse des Robust Task bei CLEF 2006 für Italienisch (aus DI NUNZIO ET AL. 2006)

Die Tabellen zeigen, dass die Systeme Systeme sehr gute und vergleichbare Ergebnisse erzielt haben. Die grafische Darstellung der Recall-Precision Kurven unterstreichen dies. Tabelle 10 zeigt die Resultate für die multi-lingualen Experimente der vier teilnehmenden Gruppen.

Tabelle 10: Ergebnisse für multi-linguale Experimente beim *Robust Task* CLEF 2006

Track	Teilnehmer Rang			
	1	2	3	4
Multilingual	jaen	daedalus	colesir	reina
MAP	27,85%	22,67%	22,63%	19,96%
GMAP	15,69%	11,04%	11,24%	13,25%
Run	ujamlrsv2	mlRSFSen2S	CoLesIRmultTst	reinaES2mtdtest

Die Ergebnisse zeigen, dass die Ähnlichkeit zwischen den Rangfolgen nach MAP und GeoAve sehr hoch sind. Lediglich an zwei Stellen ergeben sich Unterschiede, die jedoch nie die oberste Position betreffen. Bei den mehrsprachigen Ergebnissen rückt das vierte System auf Platz 2 vor und bei den mono-lingualen Ergebnissen für Italienisch tauscht das vierte System den Platz mit dem dritten.

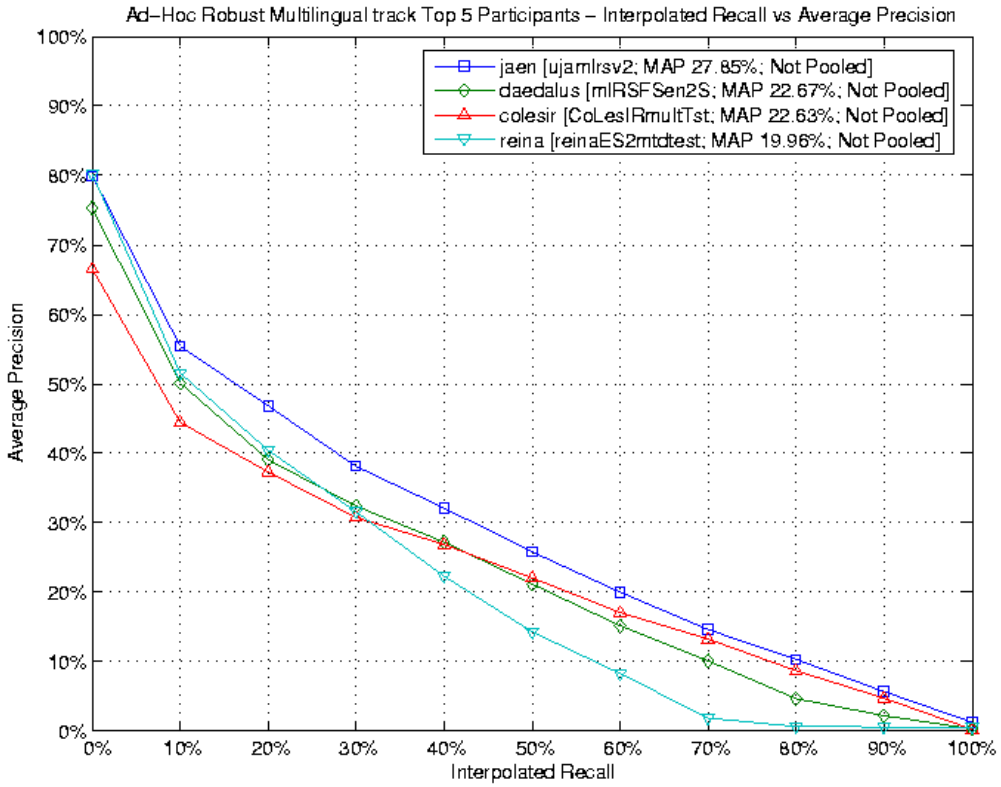

Abb. 5: Ergebnisse des Robust Task für multilinguale Experimente (aus DI NUNZIO ET AL. 2006)

Einige der Teilnehmer verließen sich auf die hohe Korrelation zwischen MAP und geoAVe und optimierten ihre Systeme nicht spezifisch für die robusten Bewertungsmßstäbe. Einige der Forschungsgruppe jedoch griffen Ansätze aus den *Robust Track* von TREC auf. Das SINAI System experimentierte mit der Expansion der Anfrage-Terme mit großen externen Korpora, die bei TREC zu den besten Ergebnissen geführt hatte (VOORHEES 2005a). SINAI nutzte dafür eine Web-Suchmaschine (MARTINEZ-SANTIAGO ET AL. 2006). Das REINA System der Universität Salamanca setzte eine Heuristik für die Identifikation der schwierigen Topics ein. Anschließend wurden für die unterschiedlich schwierigen Anfragen verschiedene Strategien zur Expansion der Anfrage-Terme eingesetzt (ZAZO ET AL. 2006). Hummingbird bewertete seine Experimente mit unterschiedlichen Evaluierungsmaßen. Besonders die Precision nach zehn Dokumenten wurde als besseres Maß als das geometrische Mittel angesehen, das die Benutzerperspektive gut widerspiegelt (TOMLINSON 2006). Das MIRACLE System wendete eine Fusion mehrerer einzelner Systeme an. Die Paramter für die Zusammenführung der individuellen Ergebnisse wurde für die Robustheit optimiert (GONI-MENOYO ET AL. 2006).

Die Analyse der Schwierigkeit der Topics wurde für die Ergebnisse fortgesetzt. Erneut zeigte sich, dass Topics nicht inhärent schwierig sind, sondern nur für bestimmte Kollektionen. Damit erweisen sich für die unterschiedlichen Zielkorpora in den einzelnen Sprachen jeweils andere Topics als sehr schwer. Zur Illustration seien hier einige Beispiele angeführt. Das Topic 64 ist das leichteste für mono-linguales und bi-linguales Retrieval mit Deutsch als Zielsprache. Für Italienisch dagegen ist es das schwerste Topic. Topic 144 führt zu den besten Ergebnissen für bi-linguales Retrieval für holländische Dokumente. Für die Zielsprache Deutsch führt es dagegen zu den schlechtesten Ergebnissen.

5 Planungen für 2007

Die neu vorgeschlagenen Maße führten zu intensiven Diskussionen beim CLEF Workshop. Die Precision nach zehn Dokumenten (TOMLINSON 2006) wurde von vielen Teilnehmern als mögliche Variante begrüßt. Auch die Anzahl der Fehlschläge eines Systems wurde als Alternative erwähnt. Dafür könnte die Anzahl der Topics unter einem gewissen Schwellenwert für die Precision gewertet werden. Hier muss allerding festgestellt werden, dass das ursprüngliche Benutzermodell von TREC von einem Benutzer ausgeht, das sehr viele Dokumente bewertet. Mit der Einführung neuer Maße wird implizit auch ein neues Benutzermodell eingeführt. Dies kann natürlich durchaus sinnvoll sein.

In CLEF 2007 sollen diese Maße eingesetzt werden. Die Datengrundlage soll verändert werden. Für Englisch und Französisch stehen Topics aus sechs Jahren zur Verfügung. Davon sollen drei Jahre für das Training und der Rest für den Test eingesetzt werden. Zusätzlich sollen Topics aus drei Jahren für Portuguieisch angewandt werden, ohne dass hier Trainingsdaten bereit stehen. Als multi-lingualer Task soll bi-linguales Retrieval von Englisch zum Französischen möglich sein.

6 Fazit und Ausblick

Ein Benutzer eines Information Retrieval Systems hat nie die Perspektive einer großangelegten Evaluierurungsstudie wie TREC oder CLEF sondern sieht immer nur die Performanz eines Systems für seine aktuelle Fragestellung. Der Robsut Task bei CLEF 2006 stellt die Benutzerperspektive in das Zentrum und evaluiert Systeme danach.

Der Robust Task stellt über die CLEF Initiative 2006 hinaus ein wertvolles Forschungsinstrumentarium zur Verfügung. Die Kollektion kann weiter benutzt werden. Auch die erzielten Ergebnisse werden weiterhin ausgewertet.

Danksagung

Der *Robust Task* wäre nicht möglich gewesen ohne die Hilfe zahlreicher Personen. Im Vorfeld und begleitend hat Ellen Voorhees (NIST, USA) bei allen Entscheidungen beraten. Ein *Robust Task Committee* hat die das Design der Aufgaben betreut. Dazu gehörten: Donna Harman (NIST, USA), Carol Peters (ISTI-CNR, Italien), Jacques Savoy (Universität Neuchâtel) und Gareth Jones (Dublin City Uuniversity). Dank der Hilfe von Giorgio di Nunzio und Nicola Ferro (Universität Padua) konnte der *Robust Task* überhaupt durchgeführt werden. Sie haben beim Task Design mitgewirkt und die CLEF Infrastruktur (das DIRECT System) angepasst. Zuletzt sei allen Teilnehmern gedankt, die Ihre Zeit in die Experimente investiert haben, um so zum Gelingen des *Robust Tasks* beizutragen.

Literaturverzeichnis

Braschler Martin (2003): CLEF 2002 - Overview of Results. In Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003, Trondheim. Berlin et al.: Springer (LNCS) Preprint http://www.clef-campaign.org.

Braschler, Martin; Peters Carol (2004): Cross-Language Evaluation Forum: Objectives, Results, Achievements. Information Retrieval. no. 7. 7-31.

Buckley, Chris; Voorhees, Ellen (2005): Retrieval System Evaluation. In: TREC: Experiment and Evaluation in Information Retrieval. Cambridge & London: MIT Press. pp. 53-75.

Buckley, Chris (2004): Why current IR engines fail. In: Proceedings of the 27th annual international conference on research and development in information retrieval (SIGIR 2004), New York: ACM Press. S. 584-585.

Cronen-Townsend, S.; Zhou, Y.; Croft, W. (2002): Predicting Query Performance. Proc Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '02) (Tampere, Finland, Aug. 11-15, 2002). ACM Press, 299-306.

Eguchi K, Kando Noriko; Kuriyama K (2002): Sensitivity of IR Systems Evaluation to Topic Difficulty. In Araujo CPS and Rodríguez MG (eds.). Proc Third International Conference on Language Resources and Evaluation (LREC) (Las Palmas de Gran Canaria, Spain, May 29-31), 585-589.

Goni-Menoyo, José; Gonzalez-Cristobal, José; Vilena-Román, Julio (2006): Report of the MIRACLE teach for the Ad-hoc track in CLEF 2006. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Harman, Donna; Voorhees, Ellen (1997): Overview of the Sixth Text REtrieval Conference. In Harman D and Voorhees E (ds.). The Sixth Text REtrieval Conference (TREC-6). NIST Special Publication, Gaithersburg, Maryland, 1997, http://trec.nist.gov/pubs/

Harman, Donna; Buckley, Chris (2004): RIA and 'Where can IR go from here?'. In: ACM SIGIR Forum, vol. 38, (2). S. 45-49 .

Kwok, K (2005): An Attempt to Identify Weakest and Strongest Queries. In: SIGIR Workshop Predicting Query Difficulty. 2005. http://www.haifa.il.ibm.com/sigir05-qp

Mandl, Thomas.; Womser-Hacker, Christa (2005): The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. Proc ACM Symposium on Applied Computing (SAC). Santa Fe, New Mexico, USA. March 13.-17. S. 1059-1064.

Mandl, Thomas (2006): Neue Entwicklungen bei den Evaluierungsinitiativen im Information Retrieval. In: Mandl, Thomas; Womser-Hacker, Christa (Hrsg.): Effektive Information Retrieval Verfahren in der Praxis: Proceedings Vierter Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2005) Hildesheim, 20.7.2005. Konstanz: Universitätsverlag [Schriften zur Informationswissenschaft 45] S. 117-128.

Martinez-Santiago, Fernando; Montejo-Ráez, Atruro; Garcia-Cumbreras, Miguel; Ureña-Lopez, Alfonso (2006): SINAI at CLEF 2006 Ad hoc Robust Multilingual Track: Query Expansion using the Google search engine. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Mothe, Josiane; Tanguy, L. (2005): Linguistic features to predict query difficulty - a case study on previous TREC campaigns. In: SIGIR Workshop Predicting Query Difficulty. http://www.haifa.il.ibm.com/sigir05-qp

Di Nunzio, Giorgio; Ferro, Nicola; Mandl, Thomas; Peters, Carol (2006): CLEF 2006: Ad Hoc Track Overview. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Peters, Carol Braschler, Martin, Gonzalo, Julio; Kluck, Michael (2003) (eds.): Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2002, Rome. Berlin et al.: Springer [Lecture Notes in Computer Science 2785]

Peters, Carol; Braschler, Martin, Gonzalo, Julio; Kluck, Michael (2004) (eds.):Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003, Trondheim. Berlin et al.: Springer [Lecture Notes in Computer Science] Preprint http://www.clef-campaign.org.

Savoy, Jaques; Abdou, Samir (2006): UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Tomlinson, Stephen (2006): Comparing the Robustness of Expansion Techniques and Retrieval Measures. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Vilares, Jesús; Oakes, Michael, Tait, John (2006): CoLesIR at CLEF 2006: Rapid Prototyping on an N-gram-based CLIR System. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Voorhees, Ellen (2005a): The TREC robust retrieval track. In: ACM SIGIR Forum 39 (1) 11-20. Voorhees, Ellen (2005a): The TREC robust retrieval track. In: ACM SIGIR Forum 39 (1) 11-20.

Voorhees, Ellen (2005b): Overview of the TREC 2005 Robust Retrieval Track. In: Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005) Gaithersburg, Maryland, November 15-18, 2005. http://trec.nist.gov/pubs/trec14/t14_proceedings.html

Womser-Hacker, Christa (1996): Das MIMOR-Modell. Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval. Universität Regensburg, Habilitationsschrift.

Zazo, Angel; Figuerola, Carlos, Berrocal, José (2006): REINA at CLEF 2006 Robust Task: Local Query Expansion Using Term Windows for Robust Retrieval. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.